Catching the Drift

When Regimes Change over Time

DISSERTATION

DER WIRTSCHAFTSWISSENSCHAFTLICHEN
FAKULTÄT
DER UNIVERSITÄT ZÜRICH

zur Erlangung der Würde
eines Doktors der Informatik

vorgelegt von
PETER VORBURGER
von
St. Margrethen (St. Gallen)

genehmigt auf Antrag von
PROF. DR. A. BERNSTEIN
PROF. DR. M. PAOLELLA

Januar 2009

Die wirtschaftswissenschaftliche Fakultät der Universität Zürich, Lehrbereich Informatik, gestattet hierdurch die Drucklegung in der vorliegenden Dissertation, ohne damit zu den darin ausgesprochenen Anschauungen Stellung zu nehmen.

Zürich, den 2. April 2008*

Der Bereichsvorsteher: Prof. Dr. Gerhard Schwabe

* *Datum des Promotionstermins*

It is not the strongest of the species that survives,
nor the most intelligent,
but rather the one most responsive to change.

attributed to Charles Darwin

Table of Contents

1

Introduction

The goal of this thesis is the identification of time-dependent relationships between different data streams[1]. The application field is the determination of dominant factors influencing exchange rates. Finance experts call such a dominant factor "regime". These factors change over time and therefore this problem has been named regime drift. The early and reliable identification of such changes is crucial for finance research.

The background of this work is the following. We were approached by the fixed income department of UBS Switzerland. UBS is a finance institute which holds a foreign exchange market share of 14.85% [3]. They are interested in an illustration of time-dependent relationships between various economic variables and the foreign exchange rate between Swiss francs and US dollars (CHF/USD). The goal is a better understanding of the market situation leading to more reliable exchange rate predictions. For example, they need to know which factor (like gold price, wages, price Brent per barrel, or money supply) is the main driver for the exchange rate development.

In contrast to most previous work in finance we do not perform any automated prediction task. We pursue a more human-centered approach by presenting the relationships between the variables in a comprehensive way. We believe that on the long run successful trading strongly depends on the domain knowledge and capabilities of the finance experts. Therefore, we enable them to focus on the decisions by providing the relevant information.

In the next section we discuss the problem of foreign exchange rates and the regime determination.

[1] This thesis is associated with the research field "data mining", which is a computer science domain. Data mining is about pattern identification and knowledge discovery in data. We assume that the reader of this thesis has basic knowledge about data mining[2].

[3] http://www.euromoneyfix.com/Article.aspx?ArticleID=1331250&PageID=3594 (December 6, 2007)

1.1 Motivation

According to the Triennial Central Bank Survey of Foreign Exchange and Derivatives Market Activity [CBS, 2007] there is a huge amount of capital involved in foreign exchange. Average daily turnover was \$3.2 trillion[4] in April 2007. The survey also mentions that the increase was much stronger than the one observed between 2001 and 2004. This shows the increasing importance of exchange rates. In addition to valuation effects the Central Bank Survey specifies the following factors for the turnover increase. *"Against the background of low levels of financial market volatility and risk aversion, market participants point to a significant expansion in the activity of investor groups including hedge funds, which was partly facilitated by substantial growth in the use of prime brokerage, and retail investors. A trend for institutional investors with a longer term investment horizon towards holding more internationally diversified portfolios might also have been a factor. A marked increase in the levels of technical trading most notably algorithmic trading is also likely to have boosted turnover in the spot market."*[CBS, 2007, p. 1]

The foreign exchange market is not centralized. The currencies are traded directly between the counterparties (often major banks). The price level is regulated by supply and demand of each participant. There are several possible influencing factors for the rates e.g. liquidity of the banks.

Basically, there are two ways of making profit with spot[5] foreign exchange. The first method is known as "*spread*". Spread is the distance between bid and ask price. This is the profit margin of the traders when they buy or sell currencies. The gain rising from the spread covers administrative and transaction cost, but the profit is marginal concerning trade between major banks.

The other, more profitable, way is to buy a large amount of one currency now and sell it later for a higher rate. This accumulation of a certain currency results only in a benefit when the foreign exchange rate evolves positively. Naturally, a profit is also made by first selling an amount of currency and then buying back the same amount later at a lower rate. Both of these trading strategies have the common assumption that the underlying estimation of future exchange rate movements is better than chance. This estimation of market movement is called *"sentiment"*. Often the market is influenced by numerous factors. When a certain factor appears to be dominant for the market movement, finance expert say that the market is under the "**regime**" of this factor.

Regimes usually change over time. We call a change in regimes a regime drift. This regime drift can be abrupt, or slow and continuous. The illustration in 1.1 shows such a regime drift on synthetic data. Variable 1 at the top is the target variable – in our case the exchange rate. The other two variables are time series which are regime candidates. At the beginning the target variable is parallel to variable 2, hence variable 2 is the dominating regime. After half of the time an abrupt change in the regimes takes place. Now, variable 2 has no more influence on variable 1 and variable 3 takes charge. So, variable 3 is the new regime.

[4] 1 trillion = 1'000'000'000'000

[5] Spot defines a contract for immediate delivery. In contrast to future trading, where a trade is set at a specified time in the future for a predefined price.

When experts are able to identify a regime drift before counterparties, they have an advantage in knowledge resulting in a superior trading strategy. The leverage effect of such an advantage is immense regarding the huge turnover in foreign exchange trade.

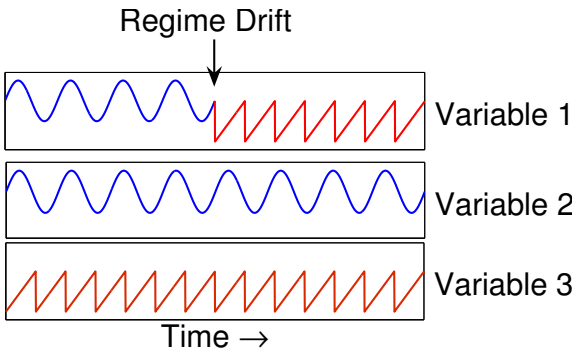


Figure 1.1: Illustration of an abrupt regime drift.

So, identifying regimes is of central importance to increase profit as well as to safeguard a currency. Nowadays, analysts and traders identify regimes by spotting parallel movements, e.g. when the exchange rate starts to move parallel to the oil price. Then, the trader knows from his domain knowledge that this might be an important factor – a regime – and starts to take the appropriate actions. In reality, the trader is faced with various difficulties during his research. One problem is the variety of factors. For example Goldman Sachs keeps track of 1.1 million financial time series [Weigend, 1997]. This is far too much. To cope with this problem, usually the observation is focused on a subset of time series. Even though there is redundant information in the overall available data, one might have omitted useful information by limiting the number of candidates. This bears the risk of omitting useful information, even though there is redundant information in the overall available data. In our case, the finance experts have chosen 86 candidates as possible influencing factors (see Section 6) which is still a large number. Here, the finance experts need assistance in the following ways.

- They need a concise, but comprehensive **overview** on the regime candidates. This includes monitoring for critical phases, i.e. hot spot detection or going on alert at the beginning of turbulent times. Also useful for the expert would be the possibility to compare the current regime situation with situations from the past so he is able to recall similar scenarios from the past.
- They need a **precise** regime representation. This includes two requirements. On the one hand accurate measurement of regime intensities and on the other hand the ability to quickly adapt to new situations. Otherwise, the expert can not rely on the regime representation or will react too late.

In this work we present an approach to this kind of problem. In our work, we mainly focus on

precise representation of the regimes. A comprehensive representation of the regimes supports the overview on all factors.

But first, we formalize the requirements in the problem definition section.

1.2 Problem Definition

According to the requirements stated in the "Motivation" Section we formulate the problem as follows:

> *For their research, finance experts require comprehensive reports on the relationships between foreign exchange rates and numerous economic variables. Since these relationships tend to change over time (regime drifts) the main challenge is to identify these relationships with high precision and detect their changes.*

To tackle the problem above we decompose the problem to well-known, solvable sub-problems. To this end, we make the following assumptions and definitions.

- **Assumptions**
  - We have to assume an open world in contrast to a closed world. A closed-world assumption states that all influences are known and appropriately represented. We cannot hold this assumption because there are manifold influences e.g. political overthrows and natural disasters which cannot be represented [Brooks, 1991].
  - Without loss of generality, we assume all relationships to be one-to-one and not many-to-one relationships.
  - We exclude self-reinforcing systems.
  - We focus on simultaneous relationships between two economic variables and neglect indicators running ahead of other variables and vice versa.
- **Definitions**
  - We define a regime as the relationship intensity of a variable with respect to another variable. The higher the relationship the higher the regime.
  - A regime drift is defined as a the change of the regime intensities with time.
  - We define relationships between two economic variables as correlations. So, we do not claim to explain or discover influences, only correlations. This has implications for the interpretation of the results (see Section "Causality", p. 16).
  - The problem definition aims at high precision. High precision is defined as being as close as possible to the real target concept. Besides a suitable correlation measure to

determine the regime intensity, this requires both adaptivity and robustness. Adaptivity is the ability to cope with new situations in short time. Robustness is insensitivity towards noise. Obviously, the goal of adaptivity and robustness is a typical trade-off, which has to be optimized to reach high precision.

- **Decomposition into Sub-Problems**

  We decompose the regime drift determination problem into two well-known sub-problems. This provides a fundament of state-of-the-art techniques to tackle our problem.

  – The first sub-problem is to find the most suitable correlation measure to determine the regime intensity. In data mining terms the correlation finding task can be associated with the field of "*feature selection and feature ranking*". Feature ranking is about determining the relevance of each feature[6] with respect to the target problem and the subsequent ranking based on a relevance score. In feature selection, a subset of the features is selected based on the ranking for further applications. In our work we only make use of the relevance determination step because this corresponds to the determination of the variables' regime intensities. Then, we hand over the regime illustrations to the finance experts who analyze the results by incorporating their (implicit, tacit) background knowledge.

  – The second sub-problem is the adaptivity towards fundamental regime changes while staying robust towards noise. The adaption to rise and fall of regimes can be viewed as a concept drift problem. A concept drift occurs when the underlying data generating mechanism changes over time. When learning a model on all of the available data, the model would be more and more inaccurate because it's underlying data set gets increasingly inconsistent. Concept drift techniques are able to determine the moment where out-dated models have to be substituted by new, more accurate models (see Section 2.2). These techniques are able to tackle the adaptivity/robustness trade-off.

The assumptions confine the scope of this work and leave space for further research and development (see "Limitations and Future Work", p. 91). The definitions and the sub-problems are discussed in detail in the "Foundations and Related Work" Chapter to provide an overview on the different solution approaches. These solutions are the foundation we build our work on.

[6] The formatted input attributes for data mining algorithms are called features. We use the name "feature" synonymous with "attribute", "variable", and "factor" since in this work we do not perform any feature construction and use the variables as features as they are.

1.3 Hypothesis

Since this work is a doctoral thesis, the goal of this work to prove/disprove the validity of a thesis. Therefore, we formulate the basic research question as a hypothesis. The hypothesis is: "It is possible to calculate time-dependent correlations between two variables with high precision."

1.4 Our Approach

We follow a bottom-up strategy to cope with drifting regimes. In the "Problem Definition" Section we decomposed the regime drift problem to two known data mining fields. Now, we combine these well-known fields to a novel solution.

There are two ways to combine these two fields: on the one hand we take the problem as a feature selection task which is subjected to concept drifts. On the other hand we can view the problem as a concept drift task where we have to perform feature selection. Figure 1.2 shows these two approaches and how they are combined based on the initial data mining fields.

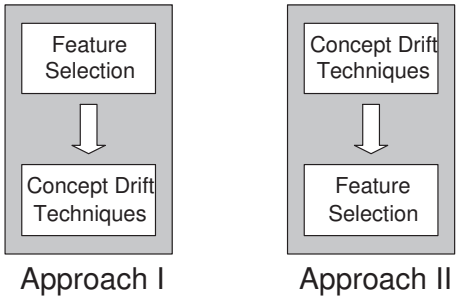


Figure 1.2: The two approaches to combine concept drift techniques and feature selection.

In the following sections, we assess both approaches on two representative implementations of these methods and decide which one to take. The selection criteria are adaptivity, robustness, and computational complexity.

The next section "Foundations and Related Work" shows that this problem has not been addressed the way we approach it. Hence, this work makes two contributions. First, we introduce the combination of the two fields "concept drift" and "feature selection". Then, we solve the real-world problem of drifting regimes on the example of exchange rates.

1.5 Organization

This work is structured as follows. First, we establish the foundation of this work which includes a review of the related work from every relevant perspective. The related work is categorized into three sections about the fields “feature extraction”, which covers the regime determination, the field concept drift, which covers the adaptivity, and an overview on data mining applications in finance.

Then, in the next two chapters we present two different approaches to cope with the regime drift problem. One approach addresses this problem from the regime determination point of view. The other approach focuses on the drift part of the regime drift problem. After that, we compare the two approaches and choose the most appropriate for our finance problem.

Subsequently, we apply the chosen approach on the finance dataset. The regime analysis is performed on 86 factors with respect to the exchange rate USD/CHF. The results are presented in a standardized way for better comparability. At the end the results and their presentation are discussed.

Next, we investigate the limitations and possible future work. We close with some concluding thoughts. For the sake of readability we have placed many calculations and experiments which are important to assert our claims in the appendix.

2

Foundations and Related Work

In this chapter we establish the foundations of our research. Therefore, we have a closer look into three research fields that are related to our work. We provide a short overview on the different fields and we discuss the relevant points for our study. To emphasize the novelty and relevance of our approach we contrast it with the work done before.

We start with "feature extraction" which contains the field of the ranking of different variables with respect to a target variable. This task contains the comparison and estimation of the relationship between variables, which is important for the determination of regime intensities. We also have a closer look on causality which is important for the interpretation of relationships between variables.

Second, we review the techniques for concept drift handling. These techniques allow adapting to the time-dependent situations.

Third, we finish with an overview on data mining in finance – our initial motivation and target application.

2.1 Feature Extraction

Feature extraction is about finding the attributes (variables, features) that represent the problem of interest in the most appropriate way.

A fundamental requirement for regime drift handling is accurate determination of regime intensities. To do this, we have to examine the relationship between a collection of candidate variables (in our case 86 micro- and macro-economic variables) and the target variable (the exchange rate CHF/USD). Figure 2.1 shows this task on a synthetic example. At the top we see the target "variable 1" and below two other variables. Feature selection – a sub-field of feature extraction – applies techniques to choose the most relevant variables for a certain problem. Here, our task is to determine the most relevant variable with respect to the target variable. The determined relevance can be viewed as the regime intensity of each variable. In the example of

Figure 2.1 the correlation has been derived by applying the Pearson correlation and the results are $r_{12} = 0.902$ and $r_{13} = 0.02$. The regime of the step curve is more intense than the regime of the zigzag curve. Therefore, in feature selection the first choice would be "variable 2" and "variable 3" would hardly be selected.

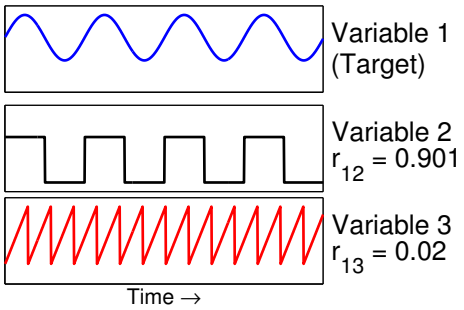


Figure 2.1: Illustration of a feature selection task.

Before focusing on feature selection we provide a short overview on some data preprocessing and feature handling techniques. All these techniques together are summarized under the name "feature extraction". The purpose of this short overview is to relate the fields "feature selection and ranking" to the other activities in "feature extraction".

Feature extraction is mostly used in predictive modeling (classification and regression). Detailed information can be found in Isabelle Guyon's book [Guyon et al., 2006]. Obviously, feature extraction is also relevant to other data mining fields like clustering [Dash and Liu, 2000, Liu and Yu, 2005]. We limit our survey on prediction problems because this will be our application field. Most of the approaches presented below can be used in other fields directly or can be extended accordingly.

2.1.1 Data Collection

The worst case for model generation is the absence of data. Often information is available, but the data acquisition is one of the most tedious and time-consuming tasks in data mining because this involves experiments and research. Experience with similar problems and background knowledge of the problem domain is of great help during data collection.

2.1.2 Data Quality

If data is available, it might still be error-prone or some information is missing. Even worse, misleading or outdated records can be part of the dataset. There are several techniques to deal with these problems. There are techniques on the data level [Dasu and Johnson, 2003] and on the algorithm level such as decision tree pruning [Quinlan, 1993] and learning of numerous models with

subsequent model combination (ensemble learning) [Opitz and Maclin, 1999]. One of the main elements of our study treats strategies against outdated records (see Section 2.2 "Concept Drift"). To estimate the impact of error-prone datasets on our algorithms, we investigate the behavior of our models under noisy conditions in Appendix A.5.

2.1.3 Feature Generation

Sometimes algorithms are performing not well enough even though the relevant information should be available in the data and the data quality is high. One reason could be an inappropriate algorithm, but in most cases the input variables do not reflect the underlying facts appropriately.

New features can be constructed by transforming or combining the original attributes. This approach is known as feature construction. It is often done by incorporating expert's background knowledge about the problem domain. For example, take a binary classification problem based on two numeric attributes. Assume, we have the background knowledge that these two attributes stand for the height and the weight of a person and the target values are "normal weight" and "abnormal weight" (underweight and overweight). Using this background knowledge we might improve the classification model by constructing a new feature like the "body mass index" $BMI = weight/height^2$. Another example is the incorporation of background knowledge about buyer behavior which can be obtained from commercial data providers based on certain buyer attributes like age, gender, or place of residence.

Other methods for feature generation work without background knowledge. They make use of an optimized mathematical representation of the input attributes. The "Principal Component Analysis" PCA [Jolliffe, 2002] is such a method for finding the most descriptive features. In detail, PCA finds the optimal linear axes transformation (rotation and stretching) by solving an Eigenvalue problem on the input attributes.

In this study, we do not transform the input variables / attributes into new features. We take the features directly from the data source without further processing, because the focus of our work is not on the feature generation level. Our focus is the relevance of the pure features.

2.1.4 Feature Selection

In some cases we are faced with the problem of having a too many possible relevant attributes. Then, the question is, which attributes are the most descriptive for our purposes? Some attributes might be irrelevant, redundant, or others contain useful information only when combined together. In short: we can't see the woods for the trees. Often an algorithm's calculation time, memory use, and performance suffers under too much input features so we need to care about this problem before feeding all possible features to the final algorithm.

The common way of dealing with a vast amount of features is choosing the most descriptive

sub-set of features for the target domain. This task of finding the most compact and informative set of features is called "feature selection" [Blum and Langley, 1997, Guyon and Elisseeff, 2003, Hall and Holmes, 2003]. In the next sections we will discuss the three feature set reduction methods.

Wrapper Methods

In machine learning a wrapper is an interpretative function that evaluates an expression to be tested and returns a value. This value allows to select the best alternative expression.

In feature selection, the wrapper is the model evaluation based on different feature combinations. The evaluation result values (e.g. the accuracy from a 10-fold cross validation) allow the identification of the best-performing model and thus, the best-performing feature combination. So, the best-performing feature combination is the feature combination to select from all features.

There are three decisions to make to perform this kind of feature selection.

First, *what is the selection criterion* to apply. Typically, the outcome of a classifier evaluation is the accuracy or the "area under the ROC curve" AUC [Provost and Fawcett, 2001] (also called the c-statistic [Cash, 1979]). These measures are the mostly used selection criteria following the rule: the higher, the better.

Second, *which algorithm* to use. Although, the wrapper approach is concerned to be a black box approach to score the feature sub-sets, the algorithm choice has some influence on the results of the final model. Maybe the algorithm used by the wrapper has less discriminative power than the subsequent learner and thus, unintentionally, omits valuable information.

Third, we have to *determine the appropriate search strategy*. Ideally, wrapper methods would make use of all possible feature combinations to determine the feature contributions (exhaustive, complete search). The state space of all possible feature combinations grows exponentially with the number of features. The number of states s grows as $s = 2^f$, where f is the number of total features. Therefore, the determination of the most relevant features using this kind of method is primarily a problem of computational complexity. As usual in computer science this problem can be represented as a search problem for which numerous solution strategies exist. Thus, most studies on wrapper methods are about finding the most efficient search strategy [Kohavi and John, 1997, Opitz, 1999]. In feature selection, there are two fundamental search procedures, the forward and backward selection. Forward selection starts from scratch and adds new variables one-by-one while evaluating the optimal search path. The backward selection does the opposite. The search starts from a model based on all variables and eliminates one-by-one. The results of both approaches can differ due to non-independent variables and different stopping points when a certain quality threshold value is reached. In other wrapper application fields also other search techniques as evolutionary search and simulated annealing are used.

Embedded Methods

Embedded methods perform variable selection during the training process of the definitive algorithm and are specific to given learning machines. In contrast to wrapper methods the embedded methods are not handling the algorithm as black box.Early examples are decision trees such as CART which have a built in mechanism to perform feature selection [Breiman et al., 1984]. More recent embedded methods guide their search for the feature sub-set by a fitness function which has to be optimized in order to reach a maximal goodness of fit and a minimal number of features [Cun et al., 1990, Weston et al., 2003].

Filter Methods

Filter methods filter out features that have little chance to be useful in the subsequent data mining steps. Wlodizislaw Duch provides a comprehensive overview on this field in [Guyon et al., 2006, pp. 89-118]. The filter is a function returning a relevance score for each feature. The estimation of a relevance score and the subsequent ranking of the features according to their scores are known as "**feature ranking**". The feature filter can be based on simple functions such as correlations, information contents, and distance measures. An example for such a correlation function is the Pearson correlation, which we will use – amongst other measures – in this study and discuss below. Relevance estimations based on information contents and distance measures are discussed in detail in [Duch et al., 2004, Hall, 1998, Dhillon et al., 2003, Forman, 2003]. More sophisticated estimations for the relevance score make use of more complex algorithms incorporating depending variables and non-linear models.

After the ranking of the features the "**feature selection**" step takes place by selecting the useful features by their relevance.

For our work the relevance scores are of central importance. One part of our initial problem definition is the regime determination. We interpret the relevance of a variable with respect to another variable as measure for regime intensities. Therefore, we make use of feature ranking techniques. The selection step after the feature ranking step might be important for the entire feature extraction process as performed in classical data mining, but for us the ranking step is sufficient. We leave the selection and other actions to the finance experts, who first have to interpret the relevance of each feature in combination with their experience. Therefore, we proceed with a closer look at some relevance estimation techniques we use in our work.

Pearson based Relevance Estimation The outcome of all relevance estimation techniques is a ranking score for each feature. To simplify the notation we call the different ranking scores "correlations". This is reasonable since a correlation describes the fact of a relevance score between two variables.

For feature ranking we use a correlation coefficient to get a measure on how strongly a vari-

able $\vec{x}$ relates to another variable $\vec{y}$. There exist numerous correlation measures [Hall, 1998]. In this study we exemplarily discuss the Pearson correlation. The Pearson correlation is defined in Equation 2.1.

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{2.1}$$

The Pearson's correlation reflects the degree of linear relationship between two variables. The output value spans a range of $\{x \in \mathbb{R} | -1 \leq x \leq 1\}$. The Cauchy-Schwarz inequality ensures that the correlation cannot exceed 1 in absolute value. A correlation of "1" stands for a high positive and "-1" for a high negative correlation. "0" stands for no correlation.

The Pearson's correlation coefficient is a parametric statistic which assumes a normal distribution (i.e., following a Gaussian distribution) of the values[1]. This assumption is reasonable for a real-world application field like finance because of the "Central Limit Theorem". The "Central Limit Theorem" states that if the sum of the variables has a finite variance, then it will be approximately normally distributed. Natural sciences like physics also use this basic assumption. Nevertheless, there exist other non-parametric correlation methods, such as Chi-square, Spearman's ρ, and Kendall's τ.

Most of the commonly used correlation functions – like the Pearson correlation – are of univariate nature, i.e. represent one-to-one relationships. Multivariate (many-to-one) aspects are considered for example in wrapper-based approaches. We will discuss such an approach below. For more theoretical background on multivariate correlations and their construction, see [Pourahmadi, 2001, Section 7.4.1].

Wrapper based Relevance Estimation For our study we used the univariate Pearson correlation and a more sophisticated multivariate correlation which we designed in a wrapper-like way. The wrapper-like correlation is constructed a little bit differently than seen in the wrapper section. In the wrapper section the best sub-set is evaluated, while the variables of a set are switched on or off. Only the overall model output was of interest. Here, in contrast, the goal is to assign a correlation value to each input variable. This is done by calculating the contribution of the variable to the model's performance. Therefore, we compare the model's accuracy with and without the variable of interest. We perform this comparison for all model combinations containing the feature

[1] If the Pearson correlation is computed on a very small number of values it is possible that all are of the same value. For this case of $y_i = \bar{y}, \forall i (i = 1, ..., n)$ the enumerator and denominator of Eq. 2.1 both tend to zero. Nevertheless, the equation can be calculated as the proof below shows. Let $\epsilon > 0$ and define $|\bar{y} - y_i| =: \epsilon$.

$$r = \lim_{\epsilon \to 0} \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot \epsilon}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} \epsilon^2}} \right) = \lim_{\epsilon \to 0} \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot \cancel{\epsilon}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot n \cancel{\epsilon^2}}} \right) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{\sqrt{n \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Because of the equation's symmetry the proof also holds for $x_i = \bar{x}, \forall i (i = 1, ..., n)$.
If both, $x_i = \bar{x}$ and $y_i = \bar{y}, \forall i (i = 1, ..., n)$ the Pearson correlation tends to 1 by applying the same proof approach.

□

of interest and take the average performance value. Equation 2.3 shows the equations for a classification problem with three input variables. For each of the three variables j the "wrapper-like correlation" r_j is calculated as follows:

$$\begin{aligned} r_1 &= 0.25 \cdot [a(m_{100}) - a(m_{000}) + \\ &\quad a(m_{110}) - a(m_{010}) + \\ &\quad a(m_{101}) - a(m_{001}) + \\ &\quad a(m_{111}) - a(m_{011})] \\ r_2 &= 0.25 \cdot [a(m_{010}) - a(m_{000}) + a(m_{110}) - a(m_{100}) + a(m_{011}) - a(m_{001}) + a(m_{111}) - a(m_{101})] \\ r_3 &= 0.25 \cdot [a(m_{001}) - a(m_{000}) + a(m_{101}) - a(m_{100}) + a(m_{011}) - a(m_{010}) + a(m_{111}) - a(m_{110})] \end{aligned} \tag{2.2}$$

In Equation 2.2 m_{101} stands for the model based on "variable 1" and "variable 3". The indices are binary and indicate whether a variable is considered (1) or not (0). The model m_{000} is the random predictor generated without any information. We have chosen the 4 most popular classifiers as underlying algorithms. These are Naïve Bayes , k-Nearest Neighbor, C4.5 decision tree (with and without pruning), and support vector machine. All these algorithms are discussed more detailed in the Appendix A.1.1. The function $a(\cdot)$ is the evaluation function. In our work the evaluation function performs a 10-fold cross-validation[2] and returns the accuracy. The correlation in 2.2 does not exceed "1" in absolute value, since the accuracy is always between 0 and 1.

The "wrapper-based" correlation function also incorporates possible dependencies between the variables in contrast to the Pearson correlation. In the following we use the naming "wrapper correlation" for the "wrapper-like" correlation.

Feature Selection on Time Series

Our target problems are financial time series. Therefore, studies on feature selection on time series are relevant for our application. The focus of these studies is the handling of vast amount of data. For example [Yoon and Yang, 2005, Yoon and Shahabi, 2006] show how to solve this problem for feature selection using different algorithms (principal component analysis, recursive feature elimination and support vector machines).

We could not find any approach considering feature selection under drifting concepts (for more information about concept drifts see next section on p. 17). Neither was a feature selection approach with forgetting capabilities found. In our study we present an approach which is able to cope with time-dependent feature relevances.

[2]For less than 10 instances we performed a leave-one-out cross-validation

2.1.5 Causality

This section is not about feature extraction itself, but nevertheless is of major importance for this field. Here, we present some rules about correlation interpretation and how to avoid typical pitfalls. For the interpretation of the correlations it is important to differentiate between correlation and causality, because *correlation does not imply causation* [Barnard, 1982]. An example is the "falling barometers" problem that tries to correlate "falling barometers" with "rain". There are four possible explanations:

1. Falling barometers are the cause of rain.
2. Some unknown third factor is actually the cause of the relationship between rain and falling barometers, e.g. a low-pressure area.
3. The correlation is coincidental. The two events occur at the same time, they have no simple relationship to each other besides the fact that they are occurring at the same time.
4. Falling barometers may be the cause of rain at the same time as rain is the cause of falling barometers (self-reinforcing system).

[Pearl, 2000] differentiates between statistical concepts and causal concepts. The examples of statistical concepts he gives are: correlation, regression, conditional independence, association, likelihood, collapsibility, risk ratio, and odds ratio. Examples of causal concepts are: influence, randomization, effect, confounding, exogeneity, ignorability, disturbance, spurious correlation, path coefficients, instrumental variables, intervention, and explanation.

Techniques to infer causal dependencies like probabilistic causality methods are based on the closed world assumption – all relevant variables are given for the domain of interest. In our target problem – the regime drifts in the finance domain of exchange rates – we cannot assume a closed world. There are always unmeasured or even unexpected confounding factors. The large number of available features (86) shows how the finance community is aware of this issue and intends to represent all influencing factors, but still there are many factors like political decisions and other real-world influences that can not be modeled [Brooks, 1991].

Here, the knowledge of the finance experts is crucial. They know how to interpret certain correlations and to find possible hidden patterns. Nevertheless, the set of all correlations supports the experts to preselect possible influencing factors for discussion, since correlations do not imply causality, but if there is no correlation the probability of causality is very low.

More information about causality and methods of inferring causal dependencies is given and assessed in [Pearl, 2000] and [Holland, 1986].

2.2 Concept Drift

In this section we show that regime drifts are related to the data mining field "concept drifts".

A concept is the underlying rule that generates the data set. In machine learning an algorithm usually learns a model, which should be as close as possible to the concept. When a concept changes (drifts) the algorithm's model needs to change too. Alexey Tsymbal provides a survey on concept drift research [Tsymbal, 2004]. He defines a concept drift as follows:

> *In the real world concepts are often not stable but change with time. Typical examples of this are weather prediction rules and customers preferences. The underlying data distribution may change as well. Often these changes make the model built on old data inconsistent with the new data, and regular updating of the model is necessary. This problem, known as concept drift, complicates the task of learning a model from data and requires special approaches, different from commonly used techniques, which treat arriving instances as equally important contributors to the final concept.*

Finance market data are subjected to external effects such as political and environmental events. Following the definition of drifting concepts above we are faced with a typical concept drift problem in finance.Harries and Horn realized the relation between financial time series and concept drifts [Harries and Horn, 1995]. They examined the movement of the stock market. Their target value for prediction was "up" or "down" for the stock market movement. They learnt a stock market model on a one month data interval and assessed their predictions during the next month. The predictions were only considered when an unseen instance is "like" the training data, i.e. only in absence of a drift. When a possible drift occurred they chose a conservative trading strategy of not taking any action. Their concept drift detection method relied only on changes in the attribute domain range. When the attribute domain range was shifted or changed in size, they assumed the occurrence of a concept drift. The predictive performance was better than chance for two of the total three month under investigation. From the weak prediction on the last month's data they inferred that there must be a concept drift between training month and target month. They could actually identify an external factor. The reason was the "Share Price Index" contract price change from $100 to $25 resulting in different trading behavior.

Although, they used a weak concept drift indicator, only the suspension of trading stocks during a drift resulted in a positive result. For our work this result emphasizes how suitable the concept drift approach is even though our targets are exchange rates, not stock markets. In contrast to the work of Michael Harries and Kim Horn we do not implement a trading strategy. We determine the relationships between the variables, visualize them, and leave the decisions to the experts by providing them the information on the factors and their intensity. Furthermore, we adaptively determine the optimal amount of history that has to be taken into consideration.

2.2.1 Types of Concept Drift

Concept drifts can occur abruptly or gradually. Examples of abrupt drifts are monetary concerns after graduation or the switch of context when a person steps into the office and disrupts [Vorburger and Bernstein, 2006a]. Examples for gradual drifts are sensor data from aging sensors or the global warming effect on the climate.

[Widmer and Kubat, 1993] differentiate between changes in the actual target concept called real concept drifts and changes in the distribution called virtual concept drifts. [Tsymbal, 2004] states that "*...from the practical point of view it is not important, what kind of concept drift occurs, real or virtual, or both. In all cases the current model needs to be changed.*" In [Vorburger and Bernstein, 2005] we developed an entropy (information content) based concept-drift method that is able to detect each of these two kinds of drifts. We also showed that virtual drifts occur when we deal with distribution sensitive measures such as accuracy, but when using distribution insensitive measures like AUC the models are not subjected to virtual drifts and thus, the models need not to be changed.

2.2.2 Handling of Concept Drift

By far most of the concept drift literature is about classification [Tsymbal, 2004]. There are also some studies about other data mining fields such as regression [Herbster and Warmuth, 1998, Herbster and Warmuth, 2001], association rule mining [Rozsypal and Kubat, 2005], and clustering [Nasraoui et al., 2003, Aggarwal et al., 2003]. The subsequent discussion is focused on classification, but holds also for other fields or can be extended accordingly.

In literature the three strategies have been introduced to keep up-to-date with the current concept: (1) instance selection, (2) instance weighting, and (3) ensemble learning. We discuss these different approaches below.

Instance Selection and Instance Weighting

In instance selection, the goal is to select the most relevant instances for the current concept [Klinkenberg and Rüping, 2003]. Instance weighting is the same, but the instances are weighted according to their individual relevance. Thus, instance selection can be formulated as a special case of instance weighting by using binary weights. [Klinkenberg and Joachims, 2000] find that instance weighting is inferior to instance selection – probably due to overfitting .

The most common procedures for concept drift handling are based on the assumption that the most recent instances are the most representative for the current concept. Thus, most procedures select the instances from a window of the most recent instances. We illustrated such a sliding window in Figure 2.2. The window size is the number of instances covered (by the window).

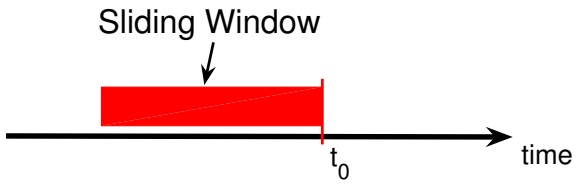


Figure 2.2: Illustration of a sliding window

Some approaches use a fixed window size [Widmer and Kubat, 1992] and others use heuristics to adjust the window size to the current situation. Most approaches adjust the window size by maximizing an evaluation measure such as the accuracy, e.g. the more advanced versions of the FLORA [Widmer and Kubat, 1993, Widmer, 1996] and FRANN system [Widmer and Kubat, 1996] and other measures like the f-measure [Kifer et al., 2004]. Other approaches compare data changes in differently sized windows. [Lazarescu et al., 2004] use three windows of different size to estimate the change in data by looking at the average vector value changes. For the concept drift estimation between differently sized windows, more sophisticated metrics e.g. entropy can be considered [Vorburger and Bernstein, 2006a].

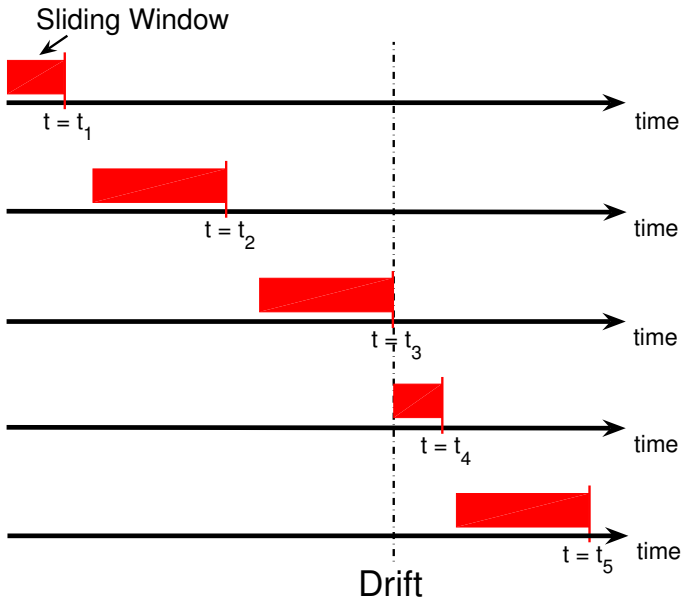


Figure 2.3: Mode of operation for a perfect window size based algorithm.

The adjustment of window size is very relevant. A too narrow window misses relevant instances, generates unstable models, and is very noise sensitive. A too wide window includes outdated instances and, therefore, the resulting models are not correct. Additionally, large window-based models are lazy towards concept changes. Figure 2.3 shows a schematic example of a window adapting its size to a concept drift. The figure represents the same problem setup of a

single abrupt concept drift on five timelines at different time steps $t_1, ..., t_5$. The topmost timeline shows the situation when the algorithm starts to learn. The second and the third timeline show the window of fixed size sweeping over the data. This is like a conventional data mining learner with constant forgetting of old instances. At t_4 the concept drift has passed by. The model built on the maximum sized window would be inconsistent and thus, outdated. Therefore, the optimal model is based on a window that is collapsed to contain only the data of the current concept. The last timeline shows the recovered model built again on the larger window size. The most popular methods of coping with concept drifts are ensembles which we discuss in the next section.

Ensemble Learning

Ensemble methods are well known from classical data mining to achieve very powerful and robust predictions [Littlestone and Warmuth, 1994, Blum and Langley, 1997]. Ensemble methods are learning algorithms that construct a set of models and then predict new data points by taking a (weighted) vote of their predictions [Dietterich, 2000]. The individual models are called experts or ensemble members and all experts together form the ensemble also known as committee. The final prediction of the ensemble depends on the decision making process which can be a (weighted) majority vote or based on other rules [Bauer and Kohavi, 1999].

The approach of ensemble methods has been successfully adapted to the concept drift domain. The ensemble's experts are models built on possibly different concepts. Continuously, the best-performing expert is chosen which the expert is built on the current concept. In the field of drifting concepts ensemble methods turn out to be accurate, flexible, and robust.

There where several different ensemble designs introduced, for example with different expert algorithms [Wang et al., 2003] or different window sizes [Kenneth, 2003, Fan, 2004]. Others extended the ensemble method with incremental updating [Chu et al., 2004] or unified the different approaches [Kolter and Maloof, 2003, Kuncheva, 2004].

In our study we make use of the ensemble method Dynamic Weighted Majority DWM introduced by [Kolter and Maloof, 2003]. We have chosen this algorithm because of its outstanding predictive performance and robustness whilst the algorithm's design is kept very simple. In Table 2.1 the pseudo-code for the DWM algorithm is listed and briefly discussed below.

The DWM algorithm maintains an ensemble of base learners, predicts using a weighted majority vote of these experts (line 11), and dynamically creates (lines 16-18) and deletes (line 14) experts in response to changes in performance (line 7 and 15). The base learners are all based on the same algorithm. In our case the Naïve Bayes , k-Nearest Neighbor, decision tree (with and without pruning), support vector machine, and linear regression. We extended the DWM algorithm from the original classification version to a regression version (see appendix A.2). We also explain how the algorithm can be extended to different evaluation measures and how the drift detecting ensemble can be decoupled from the final prediction task.

$\{\vec{x}, y\}_1^n$: training data, feature vector and target
β : factor for decreasing weights, $0 \leq \beta \prec 1$
$c \in \mathbb{N}^*$: number of classes
$\{e, w\}_1^m$: set of experts and their weights
Λ, λ: global and local predictions
$\vec{\sigma} \in \mathbb{R}^c$: sum of weighted predictions for each class
θ: threshold for deleting experts
p: period between expert removal, creation, and weight update

```
1   Dynamic Weighted Majority DWM
2
3      for i = 1, ..., n
4         σ⃗ ← 0
5         for j = 1, ..., m
6            λ = Classify (e_j, x⃗_i)
7            if (λ ≠ y_i and i mod p = 0)
8               w_j ← β w_j
9            σ_λ ← σ_λ + w_j
10        end;
11        Λ = argmax_λ σ_λ
12        if(i mod p = 0)
13           w ← NormalizeWeights(w)
14           {e, w} ← DeleteExperts(e, w, θ)
15           if(Λ ≠ y_i)
16              m ← m + 1
17              e_m ← CreateNewExpert()
18              w_m ← 1
19           end;
20        end;
21        for j = 1, ..., m
22           e_j ← Train(e_j, x⃗_i)
23        output Λ
24     end;
25  end.
```

Table 2.1: Pseudo-code for the DWM algorithm.

2.2.3 Theoretical Aspects

Research on concept drift handling is mostly of empirical nature. The reason for this might be the central characteristic of drifting concepts, the unpredictability of the next concept occurring. This does not only include the questions "will it show up abruptly or slowly?" or "what is its structure?" We do not know anything about the next concept; it could be something absolutely new. Thus, the field can hardly be covered by theory.

Nevertheless, there exist theoretical studies about defining lower and upper bounds for the window size as presented in [Helmbold and Long, 1994] and [Kuh et al., 1990].

2.2.4 Datasets for Concept Drift Assessment

In the research about concept drift handling algorithms the data sets for evaluation and benchmarking are of central importance.

Synthetic data sets are of special interest because they allow controlling the type and rate of the concepts as well as setting the noise level and adding irrelevant attributes. Unlike the real-world data sets the underlying data generator (concept) is known and allows a more profound assessment of the generated models. The most used synthetic data sets are "Stagger" [Schlimmer and Granger, 1986], the "moving hyperplane in a cube" [Hulten et al., 2001] [Wang et al., 2003], the "plane intersects a sphere" [Vorburger and Bernstein, 2006b], the SEA concept [Street and Kim, 2001], and the "moving sphere in a unit cube" [Chu et al., 2004]. All of these datasets represent two-class problems.

The "Stagger" dataset consists of discrete features and the concepts are represented by logical rules. The "moving hyperplane in a cube" defines the two-class problem by intersecting a unit cube by a plane. If an instance in the cube is on one side of the plane it belongs to class "A" and if the instance is on the other side of the plane it belongs to class "B". The SEA concept is the same, but in two dimensions. The "plane intersects a sphere" is also about the same, but instead of a cube a unit sphere is used. The symmetry of the sphere and the hyperplane rotating around the origin allows focusing on the real concept drift only – without any artifacts like distribution changes. In the "moving sphere in a unit cube" all instances are inside a unit cube. The class separation is defined by a sphere boundary inside the cube which is moved around to generate drifting concepts. The "Stagger" and the "plane intersects a sphere" datasets are discussed in detail in the next two sections.

In literature there exist also some applications on **real-world datasets** [Harries et al., 1998, Hulten et al., 2001, Street and Kim, 2001]. Unfortunately, they typically show only little concept drifts and are sometimes adapted for evaluation purposes making it difficult to assess their usefulness as a benchmark. The major problem is that the underlying concept generating the data is usually unknown.

Nevertheless, real-world datasets are very important since all concept drift handling approaches are motivated by real-world problems and are designed to be applied in the real-world. This can only be done by using real-world data. To bridge this gap we use a real-world data whose underlying data generating process is known – data taken from meteorological sensor measurements. In the meteorology dataset the drift of concept is caused by seasonal changes. We introduce this dataset after discussing synthetic datasets.

In the following, we discuss the datasets used throughout this work. There are three datasets to assess the different methods and the finance dataset for our final application. First, we present two synthetic datasets of which we know the properties of the different target concepts behind the data generation[3]. In concept drift research the predictive power on these datasets is of interest.

[3] In contrast, we focus on the time-varying correlation between the input features and the target labels.

For these two datasets we constructed different versions, where we changed the noise levels to assess the algorithm's robustness. Second, we introduce a real-world meteorology dataset where we know the period of the concept drift. lastly, the finance dataset will be discussed at the end of this study in Section "Application on Finance Data" while presenting the results.

Stagger Dataset

The "Stagger" dataset is the standard benchmark for concept drift algorithm evaluation since it has been introduced by [Schlimmer and Granger, 1986]. Schlimmer and Granger introduced this dataset together with the first algorithm dealing with this kind of problem.

Color	Shape	Size		Concept 1	Concept 2	Concept 3
green	triangle	small		false	**true**	false
green	triangle	medium		false	**true**	**true**
green	triangle	large		false	**true**	**true**
green	circle	small		false	**true**	false
green	circle	medium		false	**true**	**true**
green	circle	large		false	**true**	**true**
green	rectangle	small		false	**true**	false
green	rectangle	medium		false	**true**	**true**
green	rectangle	large		false	**true**	**true**
blue	triangle	small		false	false	false
blue	triangle	medium		false	false	**true**
blue	triangle	large		false	false	**true**
blue	circle	small		false	**true**	false
blue	circle	medium		false	**true**	**true**
blue	circle	large		false	**true**	**true**
blue	rectangle	small		false	false	false
blue	rectangle	medium		false	false	**true**
blue	rectangle	large		false	false	**true**
red	triangle	small		**true**	false	false
red	triangle	medium		false	false	**true**
red	triangle	large		false	false	**true**
red	circle	small		**true**	**true**	false
red	circle	medium		false	**true**	**true**
red	circle	large		false	**true**	**true**
red	rectangle	small		**true**	false	false
red	rectangle	medium		false	false	**true**
red	rectangle	large		false	false	**true**
Prior class distribution				88.9% (false)	55.6% (true)	33.3% (true)

Table 2.2: Stagger concepts – all combinations

The dataset consists of 120 time steps and is divided into three different concept regions each taking 40 instances. Each instance consists of three discrete feature values and a binary target. The feature values are: $color \in \{green, blue, red\}$, $shape \in \{triangle, circle, rectangle\}$, and $size \in$

$\{small, medium, large\}$. For the first 40 time steps, the target concept is $color = red \wedge size = small$. During the next 40 time steps, the target concept is $color = green \vee shape = circle$. Finally, during the last 40 time steps, the target concept is $size = medium \vee size = large$. Table 2.2 illustrates all possible feature vector combinations and the binary target labels for each of the three concepts. The table also shows the different prior class distributions in the lowest row.

For the assessment of the algorithms under noisy conditions, we introduce noise by randomly switching target labels.

Since this dataset is very short, all experiments are repeated 50 times and the results are averaged over these runs. Of course, these calculations are performed on 50 different "Stagger" datasets, all generated by the same rule presented above.

Plane Intersects Sphere Dataset

The second synthetic dataset is the "plane intersects sphere" dataset which is similar to the dataset introduced by [Wang et al., 2003] and has been applied by [Vorburger and Bernstein, 2006b]. In this dataset the three real-valued input features describe points in a three dimensional Cartesian space. Their domain is limited by a three-dimensional unit sphere (see Figure 2.4, left). A two-dimensional plane intersects the sphere through the origin, separating the instances into two-classes (see Figure 2.4, middle and right). Table 2.3 provides some sample instances of the dataset.

x-axis	y-axis	z-axis	Target Label
0.06446	0.10694	0.30232	0
-0.13922	0.38407	-0.22187	0
0.06526	-0.00899	0.17475	0
-0.00711	-0.10392	0.08745	1
-0.30884	0.09292	0.02075	1
-0.36573	0.18841	0.24551	1
-0.07086	0.12332	0.38036	0
0.34229	0.00683	0.22822	0
-0.11034	-0.00741	-0.02182	1
0.17800	-0.09793	0.25653	0
⋮	⋮	⋮	⋮

Table 2.3: Example instances of the "plane intersects sphere" dataset.

The target concept is defined by the orientation of the plane. Thus, the plane orientation in space defines a concept generating the two-class problem. A concept drift is induced by rotating the plane (see Figure 2.5). The plane's rotation axis is parallel to the third dimension such that the third feature does not affect the target concepts. Our dataset consists of 2000 time steps in total. The starting position of the plane is at -45 degrees (see Figure 2.6), lying exactly between the first two axis so the two first feature attributes have the same influence on the target concept.

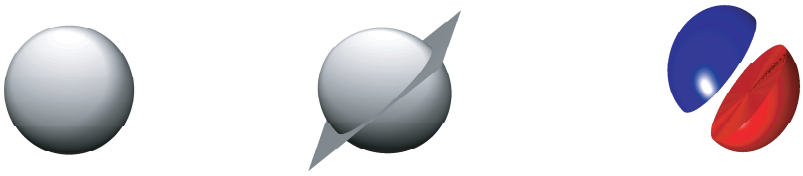


Figure 2.4: Plane separating the instances inside the sphere into two classes (blue and red).

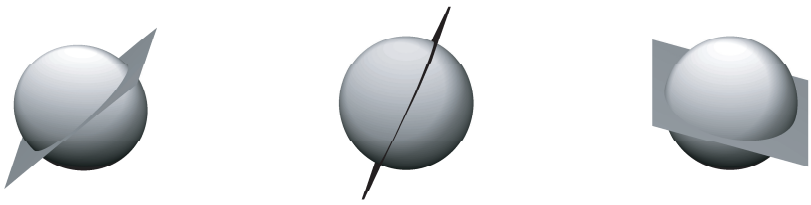


Figure 2.5: Definition of different concepts by rotating the plane.

After 500 time steps the plane flips by 180 degrees so that the new target concept is the opposite of the concept before. Then, after the next 500 instances the plane gradually rotates back to its initial starting position. This gradual shift takes 25 time steps. After a total of 1525 time steps the plane rotates gradually for 50 instances to the opposite direction and stays in this position for the remaining time steps.

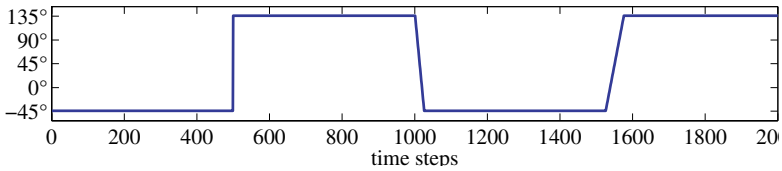


Figure 2.6: target concept definition of the "plane intersects sphere" dataset

To generate noise-prone datasets, noise is introduced like seen in the "Stagger" dataset by randomly switching target labels.

Figure 2.6 shows that this dataset allows not only assess instantaneous shifts as seen in the Stagger dataset; we can also consider gradual drifts.

Meteorology Dataset

Meteorology is a common example for continuously drifting concepts. The relationships between meteorological measurements often change during the seasons.

From the "Bundesamt für Meteorologie und Klimatologie MeteoSchweiz" we obtained a dataset containing the measurements of the "relative humidity" and the "global solar radiation" over a time period of two years (2004-2005). The data has been acquired at the Jungfraujoch – the most famous gauging station in Switzerland, located in the midst of the UNESCO World Natural Heritage site Jungfrau - Aletsch - Bietschhorn (see Figure 2.7; Source: jungfraujoch.ch, 2007). The measurement rate is 10 minutes, resulting in 144 measurements per day. We reduced the total dataset length of 105264 measurements by averaging the daily data. This results in a total of 731 instances. The reason for this reduction is the huge amount of data resulting in time-consuming calculations. This reduction does not affect the validity of our models since we are not interested in intra day predictions.

Figure 2.7: Jungfraujoch station.

Figure 2.8 shows the measurements for the relative humidity RH and the global solar radiation GSR. The global solar radiation shows the seasonal drift in a very beautiful way. The sine shape of this curve reflects the changing angle effect of the incoming solar radiation direction during the seasons. The other curve of the relative humidity does not show such distinct behavior.

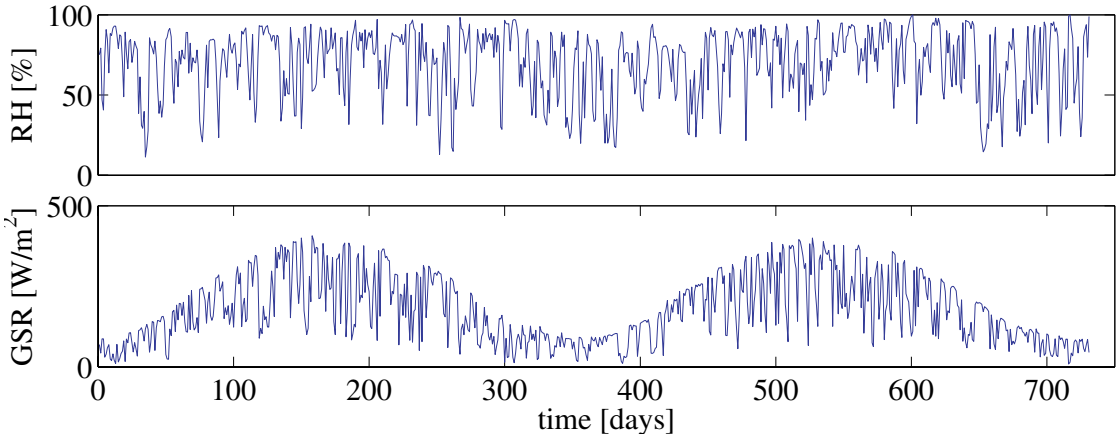


Figure 2.8: Meteorology dataset

This dataset has been chosen because it is complementary to the other datasets discussed above. First, the concept drift is continuous. Second, it is an error-prone real-world dataset with real-valued features. These properties are supposed to be similar in the finance dataset.

2.3 Data Mining and Finance

Data mining techniques play a fundamental role in financial applications [Weigend, 1997], [Nakhaeizadeh et al., 2002, Kovalerchuk and Vityaev, 2005]. The focus is on financial tasks such as forecasting stock markets and currency exchange rates. Other core tasks are the understanding and managing of financial risks, trading futures, credit rating, loan management, bank customer profiling, and money laundering analyses. In the following we provide a short overview on these fields. We also relate our work to these tasks and point out the differences.

2.3.1 Forecasting

Plenty of research [Kingdon, 1997] has been done on forecasting market behavior and financial variables. Examples are the stock markets [Harries and Horn, 1995, Rahman et al., 2002] and currency exchange rates [Walczak, 2001, Zhang and Berardi, 2001]. All studies referenced above use neural networks for prediction. For our work the study of Steven Walczak is of major interest. Therefore, we discuss it in detail. The goal of his work was to empirically find requirements on the data for financial forecasting which he investigates on the example of currency exchange rate prediction. Even though, we are not performing prediction, an investigation on data requirements is fundamental for our work. Walczak shows that models that learned an appropriate amount of historical knowledge (i.e. on a given fixed window size) outperform models using larger training sets. This has been a new fact since previous research claims that larger training sets produce better results [Zhang and Hu, 1998, Box and Jenkins, 1994, Gately and Gately, 1995]. Zhang and Hu compared predictions on window sizes of 6 and 16 years, where the predictions on 16 years performed better. In contrast Walczak showed that a window size of two years performs better. His results are supported by the Time Series Recency Effect argumentation. This effect states that constructing models with data that is closer in time to the data that is to be forecasted by the model produces a higher quality model. The conclusion is: forgetting (i.e., ignoring older data) is important. This supports our approach of introducing forgetting to our solution. The second conclusion is that the length of the window size is difficult to set and object of extensive research. All of the recent research focuses on assessing window sizes of a given fixed length. We solve this problem with our approach of applying concept drift techniques that are able to adapt the window size to the appropriate length.

Limitations of Forecasting Approaches

The intension of the approaches above is forecasting. Their goal is to have a lead in information over the other market competitors resulting in a superior trading strategy. Unfortunately, we identified two structural limitations for the forecasting approaches.

First, the information or strategy has to be unknown to the competitors, which holds only for a short period of time due to staff fluctuation and the trading strategy itself, which can be observed and – after some time – anticipated by other market participants. The result is information equilibrium. So, the application of forecasting model can be interpreted as a disturbance in a system tending to equilibrium. Nevertheless, at the beginning the profit can be considerable when using this kind of information advantage. After that, the competitors have to catch up to minimize their loss.

The equilibrium of a market with equal information for all participants is described by the efficient market hypothesis established by Eugene Fama [Fama, 1970]. The efficient market hypothesis asserts that *financial markets are "informationally efficient"*. For example, prices on traded assets, e.g., stocks, bonds, or property, already reflect all known information and therefore, are unbiased in the sense that they reflect the collective beliefs of all investors about future prospects. The efficient market hypothesis states that it is not possible to consistently outperform the market by using any information that the market already knows, except through luck.

The second limitation is that *exchange rate predictions are very difficult*. [Zhang and Berardi, 2001] for example use a traditional single keep-the-best neural networks ensemble for prediction, but they do not have a significant improvement compared to the widely used random walk model in exchange rate forecasting.

In contrast to the forecasting approaches, we do not try to forecast, we enable experts to make better predictions. Our approach is human-centered. We summarize the vast amount of information so a human expert can work more effectively. More precisely, we are the first to determine and illustrate the regimes. We are convinced that the human expert's experience in the problem of foreign exchange rate research is crucial. The experts are able to recall similar situations – even when faced in other fields. This experience acquired over years is the expert's "unique selling proposition" kept as tacit knowledge [Nonaka and Takeuchi, 1995]. In our opinion the combination of the expert's knowledge and a summarized view on important variables should result in better predictions. Even more, because of the expert-bound component, this kind of advantage over other market participants might hold for a longer time than a pure data mining-based forecasting model.

Human experts have even more advantages. Compared to computer applications, they are able to deal better with effects outside of the original boudary of attention such as political overthrows and natural disasters. Last but not least, experts might also have personal relationships to other market participants and upper level decision-makers which help to exchange knowledge or to make an arrangement during a crisis. The prime example for such a crisis is "Black Monday", where the "Dow Jones Industrial Average" dropped by 22.6% (loss of more than 500 000 000 000 dollars) on Monday, October 19, 1987. One reason was trading applications with a strategy model of blindly selling stocks as the markets fell. These feed-back effects resulted in an aggravation of the market collapse.

2.3.2 Bank Customer Profiling

Bank customer profiling as it has been part of the PKDD2000 Discovery Challenge[4] have many similarities with data mining for customer profiling in other fields [Riecken, 2000]. The goal is to get as much insight from the data to know more about the customers and to be able to take the appropriate actions.

2.3.3 Risk Management

Risk management is a field where data mining has become very important. There are different kinds of risk. Market risk is the uncertainty of future earnings due to changes in market conditions. Financial institutes also deal with credit risks. There exist numerous – already commercial – approaches to minimize those risks, e.g. JPMorgan came out with the RiskMetrics and CreditMetrics framework [Morgan Guaranty, 1994]. Another field, the country investment risk, has been investigated by [Becerra-Fernandez et al., 2002]. They predicted investing risk categories of 52 countries obtained from a Wall Street Journal survey of international experts. As input they fed their model 27 variables e.g. economic, stock market performance/risk, and regulatory efficiencies.

Another kind of risk management concerns credit ratings. [Galindo and Tamayo, 2000] provide an overview on credit risk assessment. They introduce the basic methodologies and applications. Even more, they compare different statistical and machine learning methods in this field. [Huang et al., 2004] compare a neural network and support vector machine approach on bond rating with respect to predictive and explanatory power. They also conducted a market comparative analysis on the differences of determining factors in the United States and Taiwan markets.

2.3.4 Monitoring and Auditing

Machine learning techniques are used for monitoring and auditing. One example is the simple but powerful approach using Benford's law[5] [Benford, 1938]. Hal Varian [Varian, 1972] proposed to apply Benford's law to detect possible fraud. For example, in financial accounting, the first digits which do not follow the logarithmic distribution attract attention.

Another field is the task of evaluating and forecasting banking crises. Celik and Karatepe use – again – neural network models to address this kind of problem [Celik and Karatepe, 2007]. To demostrate they examine the Turkish banking sector.

[4] http://www.cwi.nl/events/conferences/pkdd2000/ (November 5, 2007)

[5] The Benford's law states that the leading digit of real-life numbers is mostly "1". A leading digit of "2" does not occur as much as "1", but more than the number "3" and so on. The number of the leading digits follows a logarithmic distribution.

2.3.5 Financial Crime Detection

A comprehensive survey on the automated fraud detection of the last ten years is provided by [Phua et al., 2005]. In this context financial crime refers to money laundering, violative trading, and insider trading. In these fields rule pattern matching and sequence matching algorithms provide good results as the "National Association of Securities Dealers" (NASD) "Regulation Advanced Detection System" (ADS) shows [Kirkland et al., 1998, Senator, 2000]. In the ADS, for example, the pattern and sequence matcher detects predefined suspicious behaviors. In addition, new or refined patterns are identified by association rules and decision tree algorithms. Other approaches cover techniques like peer-group analysis e.g. the commercial service providers IBM and Searchspace offer[6].

For more information on money laundering we recommend the Financial Services Authority[7] (FSA) report "Review of private banks anti-money laundering systems and controls". This report summarizes the various risks originating from money laundering and the different measures (organizational and technical) against money laundering.

2.4 Conclusion

The overview of related work shows that our work is novel from all three perspectives. In feature ranking we introduce the *new approach of detecting and dealing with drifting concepts*. To concept drift research we add the *new perspective of feature assessment*. Finally, in finance research we introduce *adaptive model forgetting and accurate regime illustration*. The foundations are mostly of empirical nature. Therefore, we also mentioned the datasets, which we will use to assess our regime drift approaches.

After stating the foundations, we are prepared for the following research steps.

[6] http://www-03.ibm.com/industries/financialservices/doc/content/bin/searchspace_and_ibm_aml_brochure.pdf (October 9, 2007)

[7] http://www.fsa.gov.uk/

3

Approach I: Concept Drift on Feature Ranking

In this chapter we present the first approach to bring the two fields *feature ranking* and *concept drifts* together (see Figure 1.2). We name this approach "Approach I". In this approach we first start with given correlations. Then, we introduce a method to detect possible concept drifts based solely on the available correlation data.

There is a theoretical reason for this approach. Following the Ockham's razor argument[1], correlations are the best representation for a correlation problem.

3.1 Method Overview

First, we focus on feature ranking. To apply feature ranking on data streams, we use a sliding window technique (see Figure 2.2). The sliding window covers the last n instances of the data stream and the feature ranking score (i.e. correlation) is computed in this window like on a closed data set.

To introduce adaptivity we re-use the most-used technique applied when dealing with concept drifts: ensemble methods. Therefore, we choose a selection of experts using the same correlation determination method, but based on different window sizes (see Figure 3.1). When a regime drift occurs, we should prefer models with a shorter window-size and in absence of any drift we keep the window as large as possible to achieve more accurate results. This is a common approach in the concept drift field (see section 2.2.2).

Unfortunately, the solution is not straightforward. Classical ensemble algorithms make use of some kind of fitness function that defines which ensemble expert has to be considered. Typically, when dealing with a classification problem, the fitness function is based on parameters

[1]Named after the 14th-century English philosopher, William of Ockham. Ockham's razor suggests that the simplest hypothesis is the best. [Russell and Norvig, 2003, Mitchell, 1997]

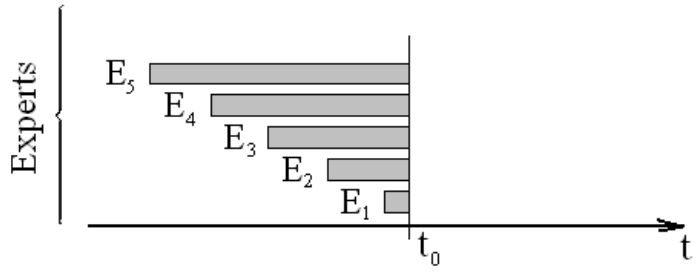


Figure 3.1: Illustration of a an ensemble made of different experts based on different sized sliding windows.

such as each expert's accuracy, the overall ensemble prediction, and possible feedbacks from past predictions. Figure 3.2 illustrates the process of classical ensemble model generation.

Figure 3.2: Process of classification under concept drift (classical)

The ensemble methods have been shown to be very powerful [Kolter and Maloof, 2003] [Wang et al., 2003], but they make use of an assumption that does not hold for this kind of approach. Their fitness function relies on the ordinal character of the input value, like the accuracy or AUC. In short, their selection criterion is based on the rule "the higher the better". In the regime drift problem, in contrast, we are dealing with correlations. *Correlations are not ordinal*. Correlations don't have to be more reliable if the correlations value is higher. A lower correlations value might be the correct one.

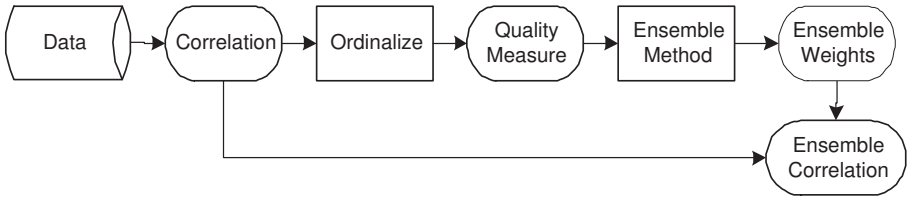


Figure 3.3: Process of concept drift method on non-ordinal measures for the case of correlations.

To reduce this new problem to a known problem we pursue the following strategy. *We convert the non-ordinal correlation values to ordinal values.* Then, having an ordinal measure, we pursue the classical process by applying well-known concept drift methods. Figure 3.3 shows the adjusted process for our approach. There are two differences to the classical process. First, we have introduced the ordinalization step. Second, we have the interaction between the ensemble-based concept drift detection and the definitive correlation determination. More details on the single process steps are provided in the next section and in the pseudo-code of an exemplary implemen-

tation in Table 3.1.

3.2 Method Formalization and Implementation

In this section we provide an exemplary implementation of the regime drift handling approach as presented in Figure 3.3. First, we provide the entire algorithm as pseudo code, followed by a detailed discussion of the ordinalization step.

Our implementation of approach I is based on an adjusted version of the Dynamic Weighted Majority (DWM) algorithm. The DWM algorithm is one of the best performing concept drift handling algorithms and has been initially designed for classification problems. Therefore, to perform the ensemble generation with the DWM algorithm we need an adjusted version of the DWM algorithm. In Appendix "DWM Algorithm for Regression Problems" on page 101 we present a DWM version for regression problems also able to handle continuous target values and not only discrete class values. The adjusted version is very similar to the original version discussed in section 2.2.2, but with a difference in the ensemble expert assessing. In the regression version the driver for considering or dropping ensemble experts is the deviation (error) between predicted and real values. Thus, this DWM version can be applied on our problem.

The link from the adjusted DWM algorithm to our application is the fitness function based on an error measure. The ordinalization step returns ordinal values which can be interpreted as error measure (see Section 3.2.1). A value of "0" stands for the highest level of correctness (no error) and the higher the value the less correct the measure is regarded. So, we use the ordinal measure direct as "error" as it would originate from a regression model evaluation.

The pseudo code in Table 3.1 formalizes an exemplary implementation of the regime drift handling method named "approach I". In this paragraph we will discuss this implementation line by line. First, we start with the initialization of the first "ensemble expert" with a window size of 1 instance and the weight of 1. This is the first range where we compute the correlation on. Then, we proceed with the outer `for` loop starting on line 4. This loop goes through the instances of the time-series occurring one-by-one. The next `for` loop on line 5 is executed for all available "ensemble experts". In this loop we first calculate the correlation r_{ij} between the input feature variable and the target variable based on the last ν_j instances. Without loss of generality, we show our algorithm for only one single input variable. More input variables would result in an additional loop, but the algorithm would stay the same since we assume the input variables to be independent of each other. Now (line 7), we convert the non-ordinal correlation value into an ordinal measure. Therefore, we pass all calculated correlations to the `Ordinalize` function (see next section 3.2.1) together with the information of the actual time step i and the expert j of interest. The return value is ξ_j, the actual ordinal measure for the expert of interest. If the ordinal measure (taken as error) exceeds the value ϑ the corresponding expert's weight is reduced by the factor β (line 9). On line 11 all weights are normalized, so the sum of all weights is 1. The

$\{x, y\}_1^n$: training data, feature x and target y
$1, ..., n$: numbering of instances, sorted by occurrence
β : factor for decreasing weights, $0 \leq \beta \prec 1$
$\{w, \nu\}_1^m$: set of experts' weights and number of instances in their sliding windows
R, r: global and local correlation values
Ξ, ξ: global and local ordinalized correlation values
θ: threshold for deleting experts
ϑ: "error" threshold

```
 1  Approach I
 2
 3     w = 1, ν = 1
 4     for i = 1, ..., n                          // Loop through time steps
 5        for j = 1, ..., m                       // Loop through experts
 6           r_ij = CalcCorr({x, y}^i_{i−ν_j+1})  // CalcCorr on window (size=ν_j)
 7           ξ_j = Ordinalize(r, i, j)
 8           if (ξ_j > ϑ)                         // Expert error > threshold
 9              w_j ← βw_j                        // Reduce expert weight
10        end;
11        w ← NormalizeWeights(w)                 // Σ_j w_j = 1
12                 R = Σ_j w_j r_ij               // Calc overall correlation
13        Ξ = Σ_j w_j ξ_j                         // Calc overall error
14        for j = 1, ..., m
15           if (w_j < θ)                         // Delete expert,
16              w_j ← 0                           //   where weight < θ
17        end;
18        if(Ξ > ϑ)                               // Create new expert
19           m ← m + 1                            //   if overall error > ϑ
20           ν_m ← 0
21           w_m ← 1
22        end;
23        for j = 1, ..., m                       // Increase window size
24                 ν_m ← ν_m + 1                  //   of all experts by 1
25                    output R                    // Return overall correlation
26     end;
27  end.
```

Table 3.1: Pseudo-code for approach I based on a modified DWM algorithm.

normalized weights allow a direct calculation of the averaged and weighted global correlation R which will be the final return value on line 25. On line 13 we estimate the global error Ξ of all ensemble experts combined together. The next `for` loop between line 14 and 17 "deletes" the single ensemble experts by setting their weights to 0 when their weight goes below the limit θ. If the estimated global error of all ensemble experts together exceeds the ϑ error threshold, a new expert is added to the ensemble (lines 19 until 21). At the end (lines 23–24) we prepare the calculation for the next instance by increasing each ensemble expert's window size by one

instance.

As we have seen in the pseudo code, the link between the ensemble based on the ordinalized values and the final correlation result is the ensemble's byproduct, the expert's weights (line 12) and window sizes. *The expert's weights describe the optimal distribution of the window size to cope with the drift.* So, when taking the weights we have the optimal window sizes to calculate the final adaptive correlation on.

Now, we focus on the central step of the ordinalization. This is the major difference between the classical concept drift and the regime drift handling mechanism.

3.2.1 Ordinalization Step

The ordinalization of a non-ordinal value is a non-trivial problem. To be able to perform this step we have to invest additional knowledge about the structure of the problem. To reach this goal we use a large number of ensemble experts having slightly different sliding window sizes. If experts with similar window sizes show considerable different outcomes, we rate them as non-reliable and, thus, we make sure they do not influence the overall outcome. Our assumption behind this approach is:

> *"If the outcome of the models for slightly varied border conditions remains similar, the outcome is more robust and thus, considered to be more reliable than if small changes of the border conditions cause large changes in the outcome"*.

This statement can be interpreted as an application of the Lyapunov stability [Lyapunov, 1992] known from chaos theory.

Our ordinalization implementation is explained in Figure 3.4. The topmost surface plot shows all possible Pearson correlations on the "Stagger" data set (for the first feature; the "Stagger" dataset has been introduced in Section 2.2.4, p. 23). The axis of abscissa stands for the time steps and the axis of ordinates stands for the sliding window size for the correlation calculation. Here, the dataset consists of 120 time steps. We calculate a set of correlations for every time step so that we have the same number of window-sizes (up to 120 instances). Obviously, the upper left corner does not contain any correlation values, since the windows cannot exceed the underlying data range. The color reflects the correlation value which can be read out from the color bar on the right of the plot. On the surface plot we can identify three different regimes. The first region is of a medium negative correlation (blue color), followed by a medium positive region (red color), and at the end a low negative correlation region (between blue to green color).

The triangle shapes in this surface plot are eye-catching. We have observed that triangle-like structures typically occur in this kind of problem. There is a reason for this kind of shape. The left diagonal border is caused by the window size of the experts and the history of data they contain. As an example the upper left region of the figure does not have any values, because the window

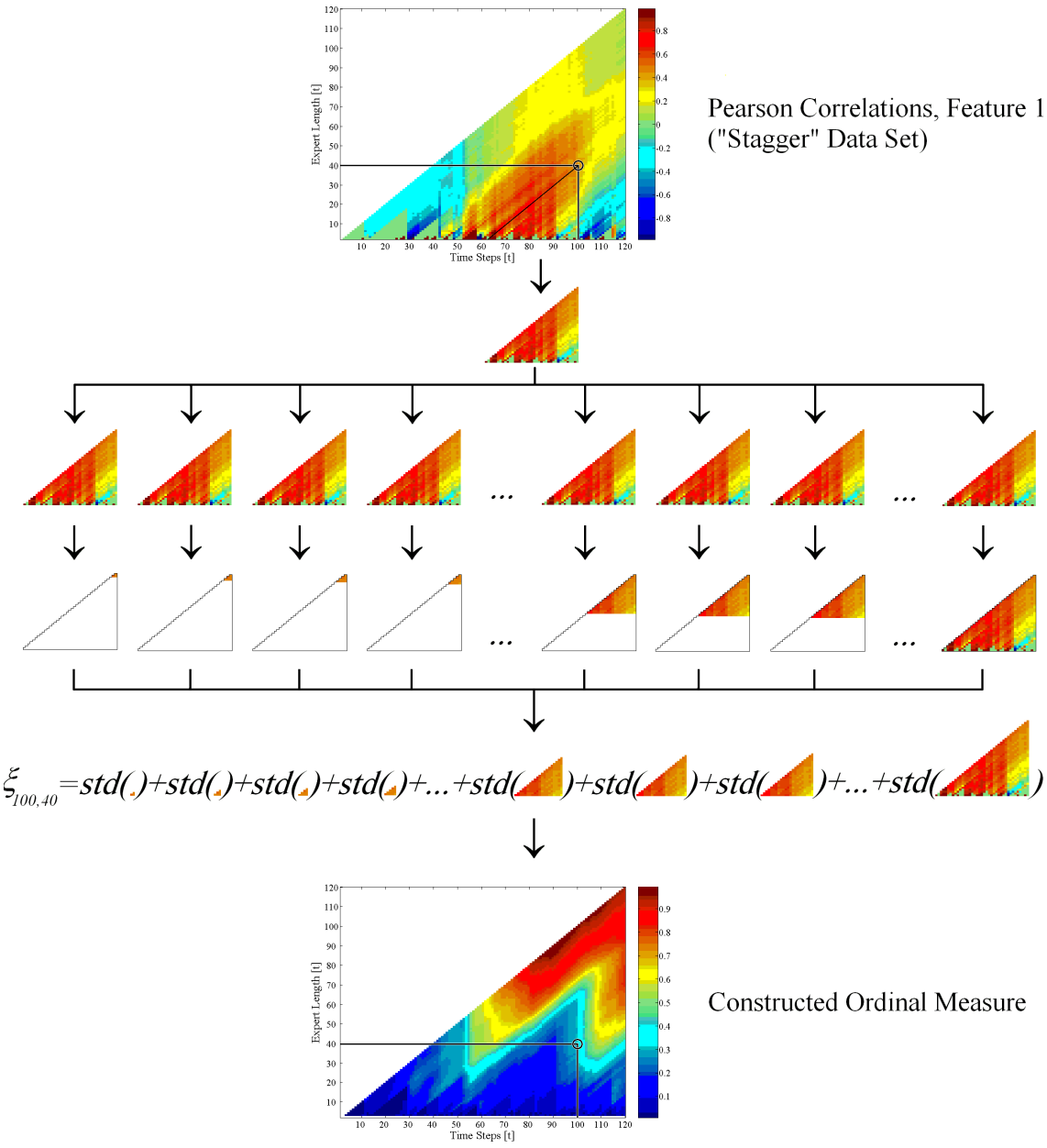


Figure 3.4: Illustration of the ordinalization of the correlation measure

sizes would be larger than the available data history. The triangle's border on the right is due to the impact of a new instance to the ensemble's expert model. If the latest instances are not consistent with the current model, the output value changes. Changes of larger window-sized experts are delayed because the older instances dominate. Smaller window-sized experts already start to recover after a few time steps and form a new triangle shape. This behavior can also be

observed for continuous drifts, where the triangle borders are smoother.

To create our ordinalization function we exploit this "triangle effect" in Figure 3.4. We explain the conversion using the correlation value at time step $t = 100$ with a window size ν of 40 instances. As shown in the middle rows of Figure 3.4, we take all possible triangles having the same top vertex at the example point and calculate the standard deviation std for all values in each triangle. Their sum results in a single value $\xi_{t=100,\nu=40}$. $\xi_{t=100,\nu=40}$ that represents our new ordinal value (see lowest surface plot[2] in Figure 3.4). When performing the ensemble-based concept drift handling we take this ordinal measure as the experts' selection criterion. In particular, the ordinal value at time step $t = 100$ with a window size of 40 instances is the criterion at time $t = 100$ for the expert of length 40.

The ordinalization function has been designed in this manner for the following reasons:

1. The standard deviation is the implementation of the stability criterion (where 0 corresponds to the most stable region).

2. The triangle-shaped data ranges exploit the "triangle effect" which typically occurs in concept drift problems.

3. The consideration of all the triangles in the formula emphasizes the region next to the expert of interest without neglecting the experts having a smaller history.

4. We use the sum-function (and not, e.g., the average) because we want to remember all "bumps".

Due to the sum-function the ordinalized values can theoretically grow without limit when the std values are non-zero. Thus, the model collapses its window size when the accumulated values exceed the acceptable ensemble expert limit in Table 3.1. In all of our experiments we have not been faced to this effect.

[2]For illustration purposes, the ξ values in surface plot have been normalized such the largest value is 1.

3.3 Results

After discussing the implementation approach I we assess it on the three different datasets. These three datasets have been presented in Section "Datasets for Concept Drift Assessment", p. 22.

3.3.1 Performance on the Stagger Dataset

Figure 3.5 summarizes the results of approach I on the "Stagger" data set. The "Stagger" dataset is the standard benchmark used in the concept drift community. It's subjected to two abrupt concept drifts.

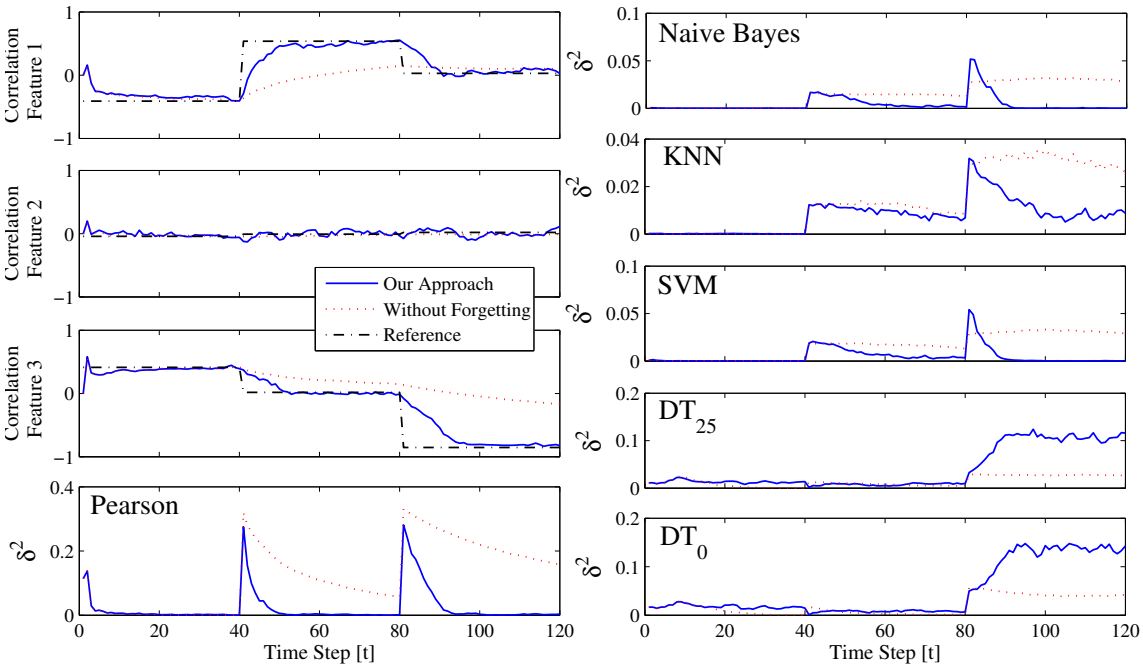


Figure 3.5: Adaptive correlations on the "Stagger" dataset for different correlation functions (Approach I)[3]. δ^2 is the average deviation from the perfect reference curves.

The left side of Figure 3.5 illustrates the results for the Pearson correlation and the right side the results of the five wrapper-based correlations[4].

The upper three plots on the left side show the Pearson correlation for each of the three features of the dataset. The dash-dot line corresponds to the perfect Pearson correlation according to the underlying data generating concept. The dotted line shows the Pearson correlation calculated

[3]The δ^2-scales are different, because we focus only on the relative comparison between the adaptive and non-adaptive case.

[4]The base algorithms are Naïve Bayes, k-nearest neighbor (KNN), support vector machine (SVM), and two versions of a decision tree (DT). DT_{25} stands for a decision tree with a confidence level setting of 25, which is the default level for C4.5 decision tree pruning. DT_0 stands for a confidence level of 0, which stands for "no pruning".

without any forgetting mechanism. Finally, the solid line shows the adaptive Pearson correlation calculated with approach I. There is a clear difference between the adaptive and the non-adaptive curves. The adaptive curves approach the reference curves over a short period in contrast to the non-adaptive curves.

The lowest plot on the left side summarizes the results of the three plots above. It shows the deviation of the adaptive and non-adaptive Pearson correlation from the perfect reference correlation. The δ^2 value is calculated by taking the squared difference between the predicted and the reference curve followed by averaging the differences over all three features. These two δ^2-curves allow drawing the same conclusions as from the three single plots above.

The plots on the right show the differences δ^2 for all five wrapper-based correlation algorithms. The support vector machine (SVM) and Naïve Bayes based algorithms show high adaptivity with our method. The k-nearest neighbor KNN based algorithm shows a bit weaker adaptivity and both decision tree based algorithms exhibit a poor performance. The reason for the poor performance of the latter two is the decision tree algorithm's performance on the rules defining the "Stagger" dataset. Decision trees usually have difficulties in predict target concepts which contains the "$\vee$" operator like in the rule "$size = medium \vee size = large$". For a small number of training examples the tree model is insufficient and therefore, our algorithm does not allow keeping those unstable ensemble experts. The non-adaptive algorithm happens to have a smaller δ^2 value because it is such inert that it remains about at the same correlation which happens to be closer to the reference.

3.3.2 Performance on the Plane Intersects Sphere Dataset

Figure 3.6 shows the results of the approach I applied on the "plane intersects sphere" dataset.

The Figure has the same layout as the Figure presented in the "Stagger" assessment. Because of the large size of this dataset we limited the maximal window size of the ensemble experts to 250 time steps. Therefore, the algorithm "without forgetting"[5] needs 250 time steps to recover.

Here, the decision tree based wrapper correlation performs much better as on the "Stagger" dataset. All other algorithms also show high adaptivity. Sometimes our approach seems to be too aggressive as the spikes in the plots indicate. First of all, the adaptive Pearson correlation for the third feature shows some distinct peaks where no peaks should appear. The curves for the other two features look fine.

3.3.3 Performance on the Meteorology Dataset

The meteorology dataset spans two years and contains two meteorological measurements, the "relative humidity" and the "global solar radiation". Figure 3.7 shows the adaptive and non-

[5] "Without forgetting" is a widely used term. Here a "limited forgetting" would be more precise.

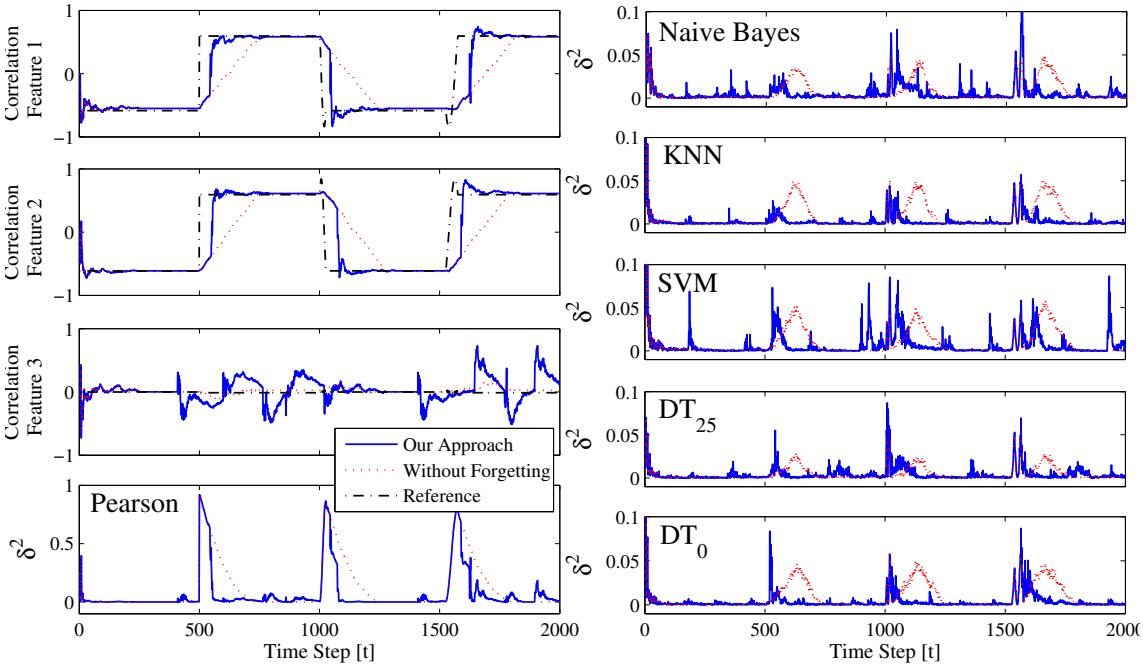


Figure 3.6: Adaptive correlations on the "plane through sphere" dataset for different functions (Approach I).

adaptive Pearson correlation between these two variables. We can identify regions in the second half of each year where the adaptive correlation differs from the non-adaptive correlation demonstrating the presence of different concepts.

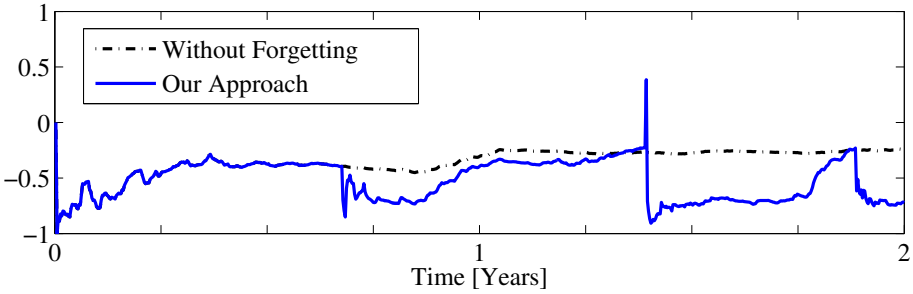


Figure 3.7: Adaptive Pearson correlation for Approach I on the meteorology dataset.

Figure 3.8 shows the ensemble weights for the non-adaptive and adaptive case. The two and three-dimensional plots on the top (a) display both the non-adaptive ensemble expert weights. The x-axis on the two-dimensional plot is the time in years. The y-axis is the memory length (window size) of each expert. The weights are exactly on the diagonal because at every time step we choose the expert with the maximal window size. While performing the calculation this stands for keeping one single expert from the beginning whose window size grows with the time steps. The coloring matches the color bar on the right. The three-dimensional plot shows exactly

the same, but the weight values are also depicted on the z-axis.

The lower plots (b) show the weights for the ensemble experts chosen by approach I. There are three reconsideration phases where the expert's memory collapses and starts to increase again.

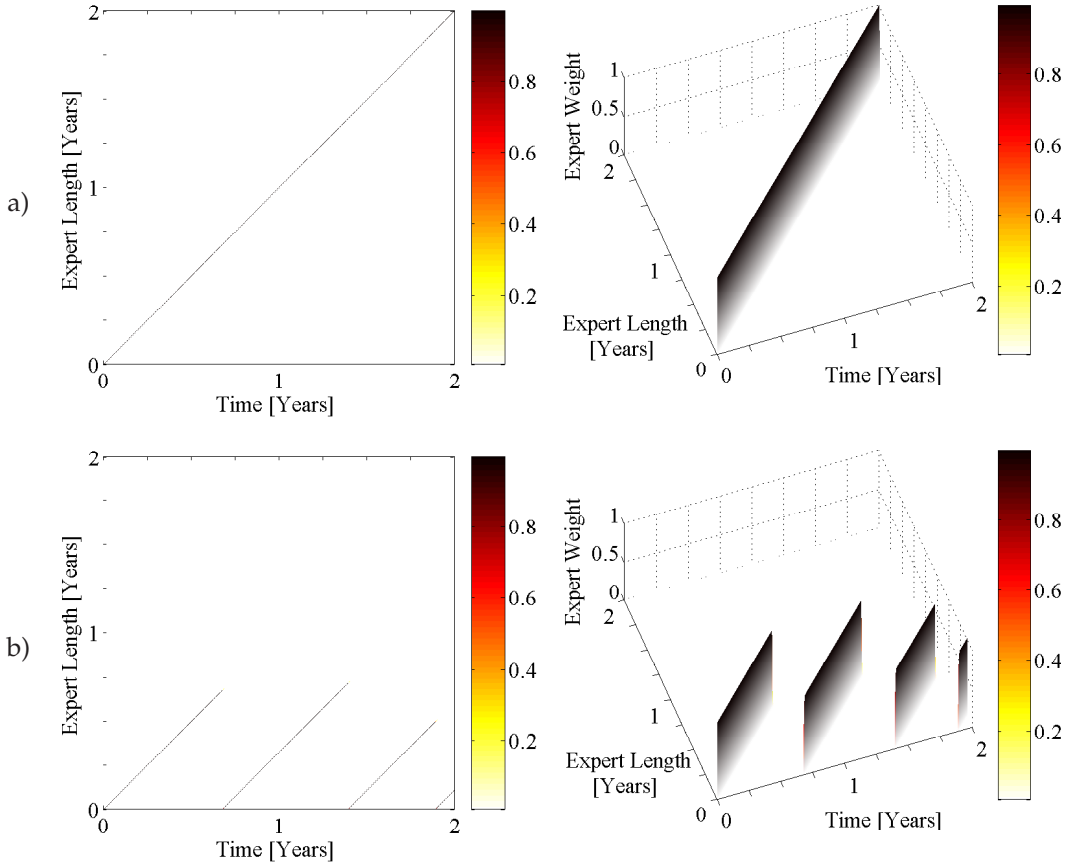


Figure 3.8: Experts weights for the meteorology dataset: a) without forgetting and b) adaptive case

3.4 Discussion

The assessment above demonstrates that approach I is able to react on concept drifts and to determine the time-varying correlations. Rarely, the method is too aggressive resulting in unwanted signal peaks as seen by the third feature in the "plane intersects sphere" dataset. The next sections are about other properties of approach I. All these outcomes will have influence on the overall comparison between the two approaches (I and II) in Section 5.

3.4.1 Computational Complexity

The computational complexity of this approach is considerably higher than for classical concept drift approaches. On the one hand we have to compute much more fine-grained expert variations and on the other hand we have the extra-step of ordinalization. That is the price we have to pay to outweigh the less information we get from the non-ordinal input values compared to classical ensemble selection methods.

The ordinalization costs are much higher than the variation costs. The computation time increases with the order of magnitude "$timesteps \cdot Experts^4$" as Figure 3.9 shows. In this example we logged the computation time during the calculation on the meteorology dataset (performed on a 3 GHz Pentium 4 machine with 1 GByte RAM). A new ensemble expert is added every time step, i.e. for each time step the maximal number of experts could reach the number of time steps elapsed. The complexity measurement has been performed by calculating all possible experts so this measurement represents the upper limit for this operation.

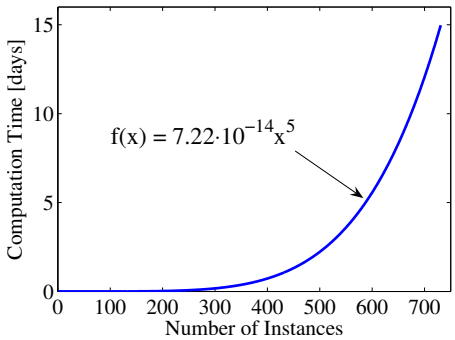


Figure 3.9: Computational complexity for approach I.

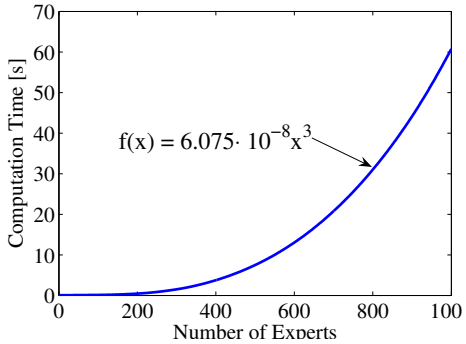


Figure 3.10: Computational complexity for the std calculation on the triangles.

The complexity of the ordinalization step is calculated as follows. The factor "$timesteps$" originates from the ensemble reconsideration done at each time step. If the reconsideration has a period larger than 1, then the factor is "$timesteps/period$" which is still an order of magnitude of "$timesteps$". The first "$Experts$" factor is the number of experts at each time step for which the ordinalization has to be calculated. The remaining factor "$Experts^3$" originates from the std com-

putation on the triangles. Figure 3.10 shows this std-computation behavior on the raw correlation values[5]. The x-axis depicts the number of experts. For each added expert the number of triangles is increased by another larger triangle as seen at the ordinalization function description in Figure 3.10. The quantity of values contained in the triangle increases quadratically. The y-axis shows the corresponding computation time. As stated above the computation time is of $O(timesteps^3)$. In both Figures the trend line fit $f(x)$ is excellent with an R-squared value[6] of $R^2 = 0.999999$ for Fig. 3.9 and $R^2 = 0.99997$ for Fig.3.10, respectively.

This behavior results in a computation time of almost 15 days for the meteorology dataset (731 time steps). If applied to the financial dataset with 6824 time steps the computation would take 2925 years.

There are alternatives to cut the computational complexity. One approach is to remodel the algorithm to an incremental version. Other approaches are e.g. limiting their maximal window size. Limiting the window size is limiting the expert's memory, like we did for the "plane intersects sphere" example. Other alternatives are more coarse-grained calculations or more efficient deviation detection algorithms than the one shown in this section. But this is beyond the scope of this work, since it requires more sophisticated methods i.e. topographic matching.

Nevertheless, we have to invest computational power in order to outweigh the lack of information provided from the non-ordinal correlation values compared to the information provided from ordinal measures.

3.4.2 Other Properties

Approach I exhibits a strong generalization properties such it can be applied to any problem, e.g. clustering, subjected to concept drifts whenever there is a ordinal or non-ordinal measure present.

In Appendix A.3 on page 103 we show how this approach performs on a classification task compared to well-known benchmarks. The outcome is of comparable performance, even though this approach does not make any use of the ordinal nature of accuracies. Hence, we can estimate that approach I is good at handling concept drifts in any field.

Appendix A.5 shows the mean δ^2 for the two datasets "Stagger" and "plane intersects sphere" depending on different noise levels from 0% to 100%. The result is a continuous loss of predictive

[6] R-squared value: An indicator from 0 to 1 that reveals how closely the estimated values for the trend line correspond to your actual data. A trend line is most reliable when its R-squared value is at or near 1. The R-squared value is also known as the coefficient of determination.

$$R^2 = 1 - \frac{SSE}{SST} \quad \text{, where}$$

Total Sums of Squares $SST = \sum_i (Y_i - \widehat{Y}_i)^2$

Error Sum of Squares $SSE = (\sum_i Y_i^2) - \frac{1}{n}(\sum_i Y_i)^2$

Y_i are the measured values

$\widehat{Y}_i$ are the values of the trend line

n is the number of data points to compare.

performance. The absence of an abrupt loss suggests a stable behavior under noise.

There exist some concept drift algorithms in literature that are able to store and remember past concepts [Widmer and Kubat, 1993]. Our approach as it is throws away previous experience without re-use. Any approach could be expanded in this direction by saving a concept library that is checked every time a new concept occurs, but this is beyond the scope of this work.

4

Approach II: Feature Ranking under Concept Drift

In this Chapter we present the second approach to bring the two fields *feature ranking* and *concept drift* together. In approach I we started from the correlations and detected the drifts on them. In contrast to approach I, in this approach we first start with the concept drift detection. Then, based on the knowledge about the drifts we apply the correlation determination methods. We name this approach "approach II".

4.1 Method Overview

Our second approach interprets the regime drift problem as a concept drift problem. From this point of view the task is to first detect the drift and then, the subsequent determination of the regime. We name the first part the "indicator" and the second part "executor".

The indicator can be one of the manifold concept drift handling algorithms as seen in Section 2.2. The choice of the indicator depends on the problem domain. The executor is one of the correlation determination methods presented in 2.1.4.

There are two fundamental questions when applying approach II. First, we need to decide how to combine the indicator and the executor so that the executor follows the indicator when facing a drift. The next Section shows an exemplary implementation of such a combination. The second fundamental question concerns the interaction between the indicator and the executor. On the one hand the indicator might not be able to model concepts which are relevant to the executor. On the other hand we are faced with dynamic effects. The indicator might be too fast or too lazy compared to the executor which is not able to catch up with the drift or drops valuable information too early. This research question is addressed at the end of the next section.

4.2 Method Formalization and Implementation

In this section we present an exemplary implementation of approach II. This implementation is of general nature so it can be extended with other indicators and executors. The only restriction is the approach of combining the indicator with the executor. For the combination we assume that the indicator's concept drift handling mechanism uses window-based forgetting (see Figure 2.3). Then, we pass the information about the windows to the executor. The executor determines the correlation on these windows and the correlation should exhibit the same adaptivity as the indicator. In Figure 4.1 illustrates the regime determination process under drifting concepts.

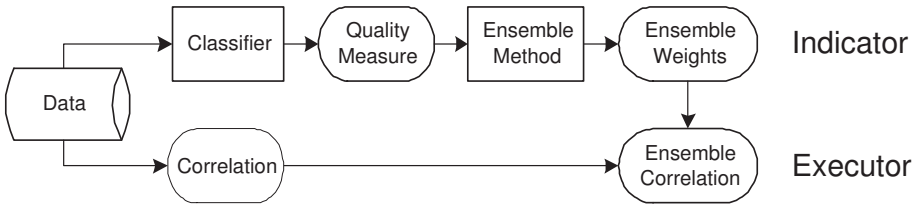


Figure 4.1: Process of calculating adaptive correlations using an external concept drift indicator

In this illustration we identify the concept drifts in a data stream by a classification (or regression) based ensemble algorithm. Therefore, we interpret the two variables[1] of interest as a prediction task dataset. One variable is taken as target, the other as input feature. In our study, we apply the concept drift detection ensemble algorithm Dynamic Weighted Majority DWM on the data stream. As byproduct we obtain the assigned ensemble expert weights from the algorithm. These weights stand for the preferred experts, i.e. for the appropriate window sizes for this kind of problem. The next step is to calculate the correlations on the same window sizes as they have been selected before by the DWM.

The pseudo-code in Table 4.1 shows the algorithm in detail. On line 3 in the left column we start with an ensemble without experts, i.e. only a random model. We proceed with the outer `for` loop starting on line 4. This loop stands for the instances appearing one after another. Then, we turn to the indicator part. From the `CalcEnsemble` method we obtain all ensemble experts from which we get each window size ν (line 7) and weight w (line 8). The `CalcEnsemble` method is listed on the right and is the DWM algorithm (Appendix A.2). After having determined the window sizes we calculate all correlations on each window (line 11). Then, we normalize the experts' weights in order to calculate the overall correlation R on line 14. R is the average of all correlations r_j with respect to the weights of the windows. Hence, the overall correlation should have the same adaptive properties as the indicator. On line 15 the value R is returned as result.

[1] We assume one-to-one relationships as mentioned in the problem definition section.

$\{x, y\}_1^n$: training data, feature x and target y
$1, ..., n$: numbering of instances, sorted by occurrence
β : factor for decreasing weights, $0 \leq \beta \prec 1$
$c \in \mathbb{N}^*$: number of classes
$\{e\}_1^m$: set of experts, their weights w, and their number of instances in their sliding windows ν
Λ, λ: global and local predictions
$\vec{\sigma} \in \mathbb{R}^c$: sum of weighted predictions for each class
θ: threshold for deleting experts
R, r: global and local correlation values

```
1  Approach II
2     e = ∅
3     for i = 1, ..., n                          // Loop through time steps
4        // Indicator part for concept drift identification
5        e ← CalcEnsemble(e, x_i, y_i)           // Calculate all experts
6        w ← GetWeights(e)                       // Get weights from experts
7        ν ← GetWindowSizes(e)                   // Get windows from experts
8        // Executor part for regime determination
9        for j = 1, ..., m                       // Loop through experts
10          r_j ← Corr({x}_{i−ν_j+1}^i, {y}_{i−ν_j+1}^i) // Correlation on each window
11       end;
12       w ← NormalizeWeights(w)                 // Σ_j w_j = 1
13       R = Σ_j w_j · r_j                       // Calc overall correlation
14       output R                                // Return overall correlation
15    end;
16 end.
```

```
1  CalcEnsemble (e, x_i, y_i) // CalcEnsemble method used in Approach II
2     σ⃗ ← 0                          // Set initial weight to 0
3     for j = 1, ..., m              // Loop through experts
4        λ = Predict(e_j, x⃗_i)       // Expert predictions
5        if (λ ≠ y_i)                // Lower weights, if
6           w_j ← βw_j               //   local prediction is false
7        σ_λ ← σ_λ + w_j             // Increase predicted class weights
8     end;
9     Λ = argmax_λ σ_λ               // Choose dominant class prediction
10    w ← NormalizeWeights(w)        // Σ_j w_j = 1
11    {e, w} ← DeleteExperts(e, w, θ) // Delete experts below θ
12    if(Λ ≠ y_i)
13       m ← m + 1                   // Create new expert, if
14       e_m ← CreateNewExpert()     //   ensemble prediction is false
15       w_m ← 1                     // Assign weight to new expert
16    end;
17    for j = 1, ..., m              // Loop through experts
18       e_j ← Train(e_j, x⃗_i)       // Include latest instance in model
19    output {e, w}                  // Return set of experts
20 end.
```

Table 4.1: Pseudo-code for approach II and the example DWM algorithm as concept drift indicator.

The crucial question when applying this approach is: what happens if the indicator has a different adaptivity than needed by the executor? For example, the indicator is too adaptive for the executor and the resulting correlation is not yet stable. This has been one reason for approach I, where we argued with Ockham's razor and thus, used the same basic values as indicator and executor.

Hence, the fundamental question behind this approach is whether a concept drift indicator based on a different algorithm is able to take the suitable actions to cope with the changing regime. We empirically investigated this on the two synthetic datasets both for a classification and a regime drift problem in Appendix A.4.2 on page 106. The results show that this application of different indicators and executors in approach II is reasonable and thus, we can proceed.

4.3 Results

In this section we assess the approach II in the same way as we did for approach I. But first, we need to define the indicator/executor mapping applied in the following calculations[2]. To combine similar indicating and executing algorithms we defined the following mapping in Table 4.2.

Indicator	⟶	Executor
Naïve Bayes classifier	⟶	Naïve Bayes based wrapper method
KNN classifier	⟶	KNN based wrapper method
SVM Classifier	⟶	SVM based wrapper method
Decision Tree classifier	⟶	Decision Tree based wrapper method
Linear regression	⟶	Pearson correlation

Table 4.2: Mapping of the indicating and executing algorithms (Approach II)

The reason for combining the linear regression and the Pearson correlation is their close relation. The Pearson correlation stands for the linear relationship between two variables (see page 13). Naturally, we mapped the decision trees with different confidence levels[3] to the corresponding wrapper methods.

4.3.1 Performance on the Stagger Dataset

The results of this approach on the "Stagger" dataset are about the same as the results seen for approach I (see Figure 4.2, p. 50). The most distinct difference is the decision tree based regime determination. In this case the correlation shows a high deviance right after the second drift like seen in approach I, but in contrast to approach I it is able to recover and tends to the correct correlation values. The cause of this behavior is the indicator DWM algorithm on the decision tree classifier which is more tolerant to unstable models than the DWM algorithm based on the ordinal measure. So, the DWM algorithm keeps some ensemble experts with low weights and allows them to recover when the expert predictions improve with larger window sizes.

4.3.2 Performance on the Plane Intersects Sphere Dataset

Figure 4.3 illustrates the results of approach II on the "plane through sphere" dataset. The adaptive Pearson correlations are performing very well for all three features in contrast to the approach I where the third feature was not properly represented (see Figure 4.3, p. 50). The performance of the adaptive wrapper-based correlation values is affected by the aggressive nature of the algorithm during the drifts resulting in high adaptivity, but also in numerous spikes. The number of

[2] See Section A.4 for all mapping combinations and outcomes.
[3] The confidence level is the threshold for decision tree pruning, see Section A.4.

spikes in the drifting regions is higher than seen at approach I. The conclusion is that the overall performance is about the same for both approaches.

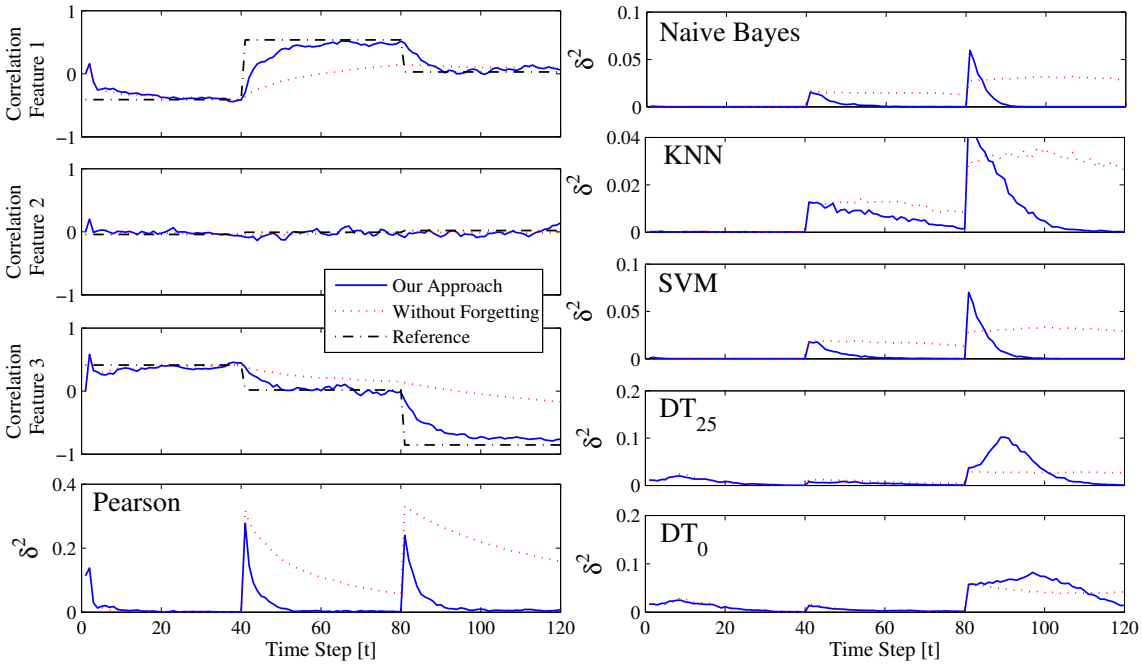


Figure 4.2: Adaptive correlations on the "Stagger" dataset for different correlation functions (Approach II).

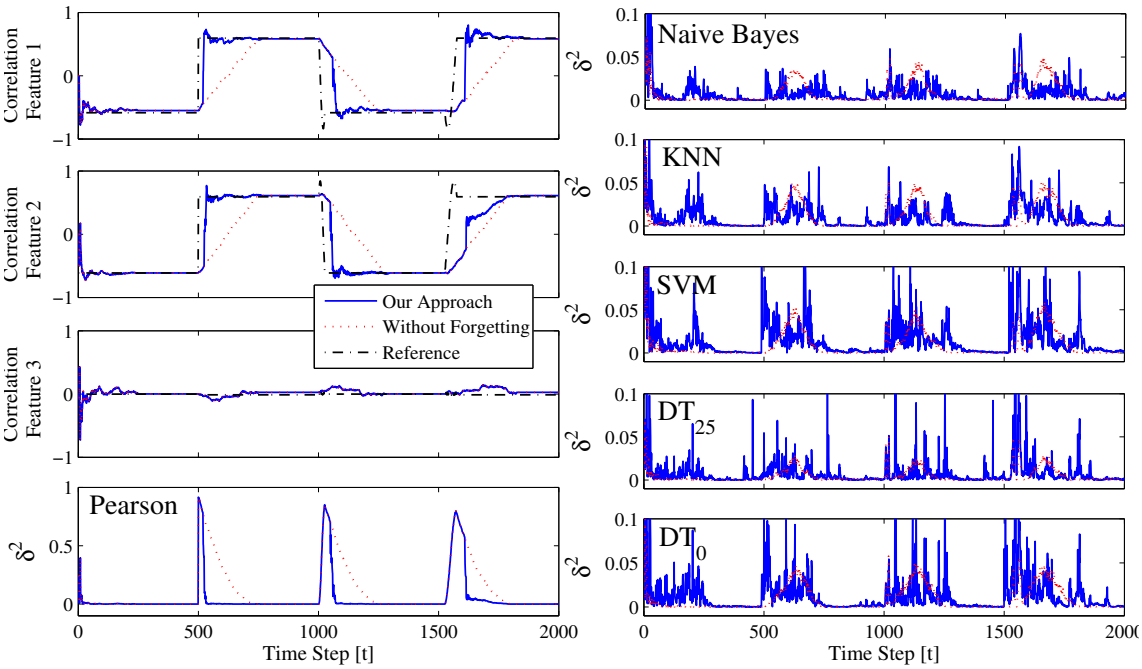


Figure 4.3: Adaptive correlations on the "plane through sphere" dataset for different functions (Approach II).

4.3.3 Performance on the Meteorology Dataset

Figure 4.4 shows the adaptive and non-adaptive Pearson correlation curves of Approach II on the meteorology dataset. There are two regions where the adaptive Pearson correlation differs from the non-adaptive curve. These regions span a period between summer and fall for both years.

The calculation using approach I in Figure 4.5 shows a similar curve with two differences [4]. First, the reaction time is faster for approach II. Second, for the last season of the second year the correlations are different. Looking at the overall shape of the curve and knowing the periodic concept behind the dataset suggests preferring the approach II solution.

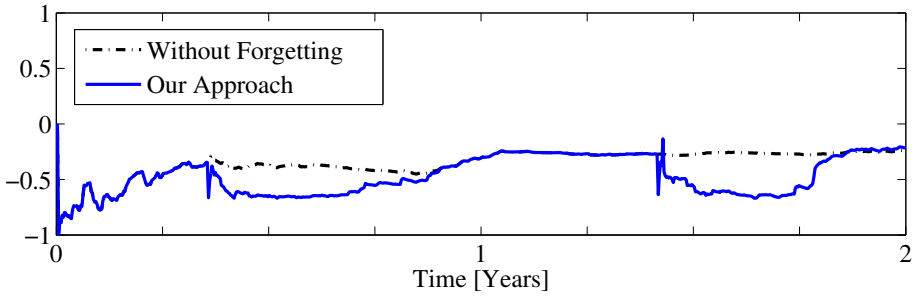


Figure 4.4: Adaptive Pearson correlation for Approach II on the meteorology dataset.

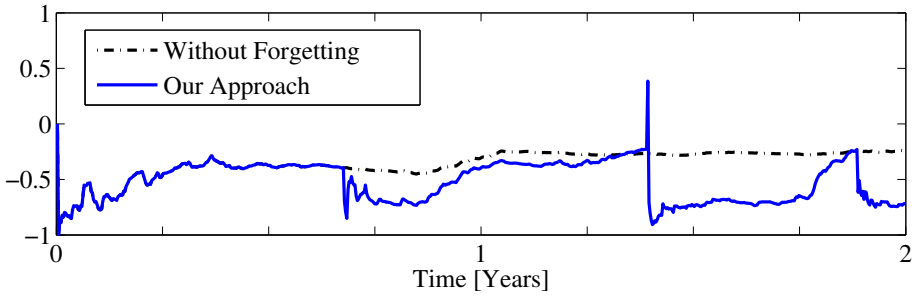


Figure 4.5: Adaptive Pearson correlation for Approach I on the meteorology dataset. (Repeated illustration to ease the comparison.)

Figure 4.6 and Figure 3.8 allow the comparison of the weights of approach II and approach I. The weight plots of approach II in Figure 4.6 provide deep insight in the meteorology dataset. The division into the seasonal changes can be recognized at first sight. The most striking facts are the two alternating regimes. The first regime almost vanishes during the other regime's period followed by recovery until the next regime rises.

[4] We repeated the illustration of Figure 4.5 to simplify the comparison.

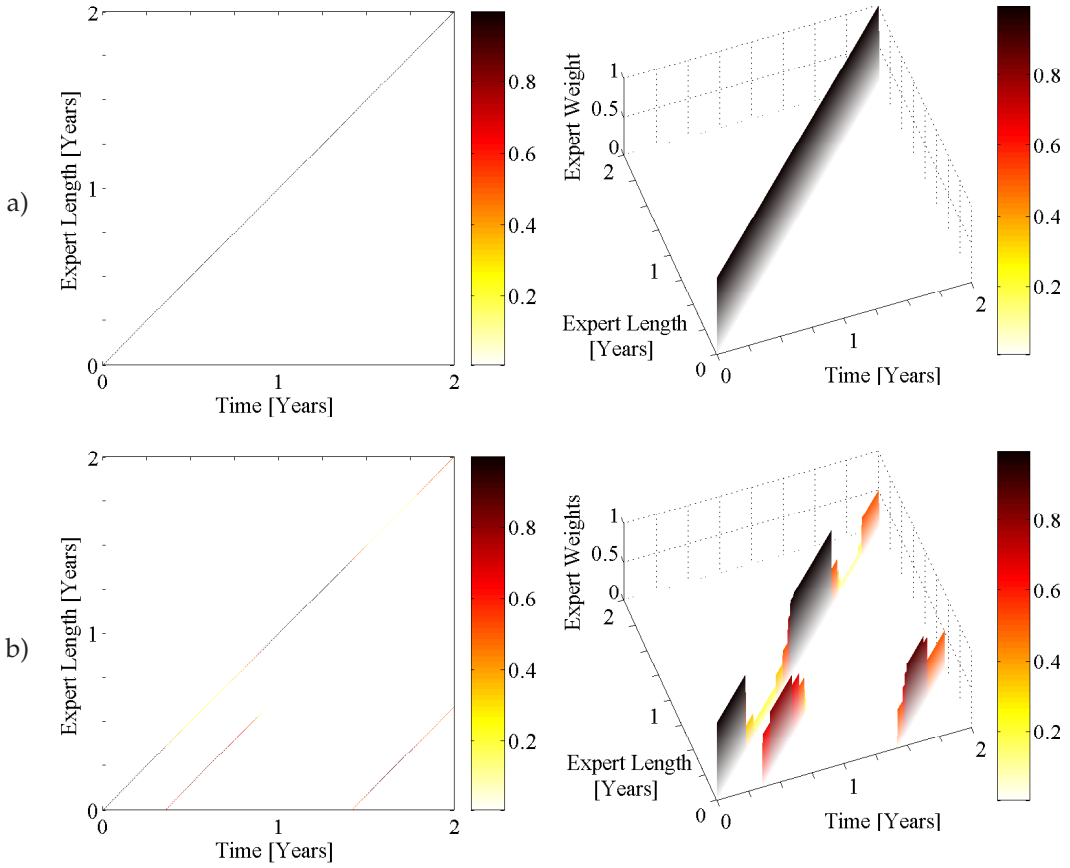


Figure 4.6: Experts weights for the meteorology dataset: a) without forgetting and b) adaptive case

4.4 Discussion

4.4.1 Review of the Results

The assessment of approach II on the three datasets above shows that this approach is suitable to cope with the regime drift problem. It is even able to handle the difficult situation of the decision trees on the stagger dataset. The adaptive Pearson correlation in the meteorology dataset reveals the seasonal changes behind the data. In contrast to those striking aspects the approach lacks some robustness for the wrapper-based correlations on the "plane intersects sphere" datasets (see 5.1 in the overall discussion, p. 56).

4.4.2 Computational Complexity

The computational complexity for the indicator consists of the computation of the base-algorithm and the DWM ensemble method. The DWM algorithm's computational costs can be neglected compared to the cost of building the expert models. Also the executing correlation determination step only depends on the computation costs of the chosen algorithm's properties.

4.4.3 Other Properties

We also assessed this approach on the two synthetic datasets under the influence of different noise levels up to 100%. The result is a continuous loss of predictive performance (see Appendix A.5). The behavior is similar to the behavior of approach I with the difference of better performance for the decision tree based algorithms on the "Stagger" dataset.

On the special prerequisite for the application of this approach is a dataset which we can express as a classification or regression problem.

Figure 4.6 shows that past concept models have been re-activated, but we do not intentionally store a collection of past concept descriptions.

5

Comparing Approach I with Approach II

In this chapter we compare the two approaches presented in the last two Chapters. Then we pick the most suitable one for the finance domain.

The comparison is conducted with respect to the following criteria. The most important is the performance in terms of adaptivity and robustness. This criterion ensures the high precision of the results. The next criterion is the computational complexity. After that, we have a look at other factors which may have some influence on the decision.

5.1 Criterion 1: Adaptivity and Robustness

The central criteria in the problem definition are adaptivity and robustness. Adaptivity is the ability to cope with new situations in short time. Robustness is insensitivity towards noise. Adaptivity and robustness are typical trade-off antagonists. Here, we look for the approach handling this trade-off best.

First, we have a look at synthetic dataset assessment. Table 5.1 shows the averaged deviations of both approaches from the perfect benchmark $\mathrm{avg}(\delta^2)$ for both synthetic datasets. The lower the deviation, the better the approach. The light gray background denotes the better value.

On the "Stagger" dataset both approaches (compare Figure 3.5, p.38 and 4.2, p.50) perform similar – except for the decision tree wrapper based correlation where approach I is not able to recover like approach II. Table 5.1 shows that approach II is slightly superior to approach I, especially for the decision tree based wrapper correlations.

On the "plane intersects sphere" dataset both approaches exhibit advantages and disadvantages. Approach I is performing better than approach II for the wrapper-based correlation computations (see Table 5.1). On the other hand (at the Pearson correlation) approach I shows too

aggressive adaptivity when applied on the irrelevant third attribute (see Figure 3.6, p.50). This also affects the average deviation in Table 5.1.

The behavior of approach II is complementary to the behavior of approach I. Approach II is more robust when dealing with the third feature in the Pearson case, but is too adaptive when applied to the wrapper-based correlations. This is a typical example of the adaptivity / robustness trade-off. Whenever we aim at high adaptivity there is a limit where robustness begins to suffer. Hence, we classify the behavior of both approaches as similar. The behavior under noise influence is also similar for both approaches (see Appendix A.5, p.109).

Dataset	Correlation-Measure	Approach I $\mathrm{avg}(\delta^2)$	Approach II $\mathrm{avg}(\delta^2)$
"STAGGER"	Pearson	0.0227	0.0197
	Wrapper (NB)	0.0042	0.0027
	Wrapper (KNN)	0.0074	0.0061
	Wrapper (SVM)	0.0047	0.0040
	Wrapper (DT_0)	0.0495	0.0230
	Wrapper (DT_{25})	0.0398	0.0161
Plane through sphere	Pearson	0.0736	0.0535
	Wrapper (NB)	0.0053	0.0076
	Wrapper (KNN)	0.0031	0.0098
	Wrapper (SVM)	0.0065	0.0144
	Wrapper (DT_0)	0.0036	0.0126
	Wrapper (DT_{25})	0.0041	0.0078

Table 5.1: Comparison of both approaches by the average δ^2.

Second, we look at the meteorology dataset. Both approaches are able to identfy the seasonal drift. The adaptive Pearson correlation looks more accurate when generated by approach II. It nicely reflects the concept cycle for all seasons and reacts faster on the changes (higher adaptivity).

Our conclusion is that the overall performance of the two approaches is comparable, but with an advantage for approach II.

5.2 Criterion 2: Computational Complexity

Computational complexity is important since the target finance application data spans a range of decades and there are almost one hundred variables to examine. Even though an update period of one day is sufficient, the computational complexity can be a limiting factor. Since both approaches base on the same algorithms and the same ensemble selection method the major difference in terms of computational complexity is the ordinalization step in approach I. The computational costs are very high for this step as explained in detail in Section 3.4.1 on page 42. Therefore, approach II is preferable from this point of view.

5.3 Decision

Based on the slightly better outcome of criterion 1 and the advantage on computational complexity we decide to go on with approach II and apply it on the finance data in the next chapter.

The generalization property of approach I to other data mining fields is not a deciding factor for application on the finance domain.

6

Application on Finance Data

In this Chapter we apply our research results to a real-world problem. The real-world problem is the exchange rate regime drift visualization task which has been the motivation for our research in this field. For the calculations we decided to apply approach II based on the outcome of the comparison chapter before.

In particular, we use a concept drift indicator based on the Dynamic Weighted Majority DWM ensemble. In our case the underlying algorithm of the DWM is the linear regression. We have chosen linear regression because of the continuous value range when dealing with most finance variables. The correlation determination is performed by the Pearson correlation (executor). The Pearson correlation is suitable for use in combination with a linear regression indicator since both are linear methods. Even more, the Pearson correlation is fast in computation and widely used.

The structure of this chapter is the following. First we provide an overview on all finance variables used in our application. Then, we explain the presentation of the results which are presented in the subsequent sections. For purposes of clarity, the presentation of the results is accompanied by a short definition of the variables. At the end we close with a recapitulation and discussion of the results.

6.1 Dataset

We investigated 77 variables (features) with respect to the foreign exchange rate between Swiss franc and the dollar (FX CHF/USD). We have chosen the Swiss franc as target on request of the finance experts.

All raw data presented in this section are available at Bloomberg and/or the Swiss National Bank SNB. We categorized the variables into three groups as the listing in Table 6.1 below shows:

Market-specific variables

- Foreign Exchange (spot) (p. 65)
- Currency Swap (p. 67)
- Commodities (p. 68)
- Forward Foreign Exchange Rate (p. 66)
- Key Interest Rates (p. 69)
- LIBOR (p. 70)
- Treasury Bonds (p. 71)
- Forward Rate Agreements (p. 73)
- Futures Short-Term Interest (p. 74)
- Stock Exchange (p. 75)

Macro-economic variables

- Gross Domestic Product (p. 77)
- Money Supply (p. 79)
- Consumer Price Index (p. 81)
- Producer Price Index (p. 81)
- Industrial Production Index (p. 83)
- Purchasing Managers Index (p. 84)
- Unemployment Rate (p. 85)
- Wages (p. 86)

Soft Factors

- Consumer Confidence Index (p. 87)

Table 6.1: Overview on the finance data.

Without loss of generality, we limited our study on the economic regions United States of America (US), European Union (EU), and Switzerland (CH). These three regions are of high interest for the Swiss market. Furthermore, the data is available and widely used.

The dataset spans a time range from January 2nd, 1980 to February 28th, 2006. Even though some of the variables (macro-economic variables from the US) have been available for more than eighty years, this time range has been chosen to best fit the average time range of all variables of interest.

The dataset consists of 6824 instances each representing a working day. Weekends and high days are not included. Non-daily variables, such as quarterly published variables, are transformed into daily variables by repetition of the last known value[1]. Besides this, the data is not preprocessed at all. Even outliers have not been corrected to keep the real-world data setup.

[1]Some variables such as the GDP are sometimes published with a delay, re-estimated and revised, but without any change history. So, we were not always able to retrace the momentary knowledge at publication time.

6.2 Presentation of the Results

To assist finance experts we provide an interpretation instruction together with some background information about the origin of the presented curves. Therefore, the results are always illustrated in the same way in a figure block as shown in Figure 6.1.

The figure block contains five sub figures. The three sub figures on the right side contain the two variables to compare and the resulting curve reflecting the comparison in terms of correlation. The two sub figures on the left provide some background information about the problem. These two plots might give some more insight to finance experts - more than one single result curve can do.

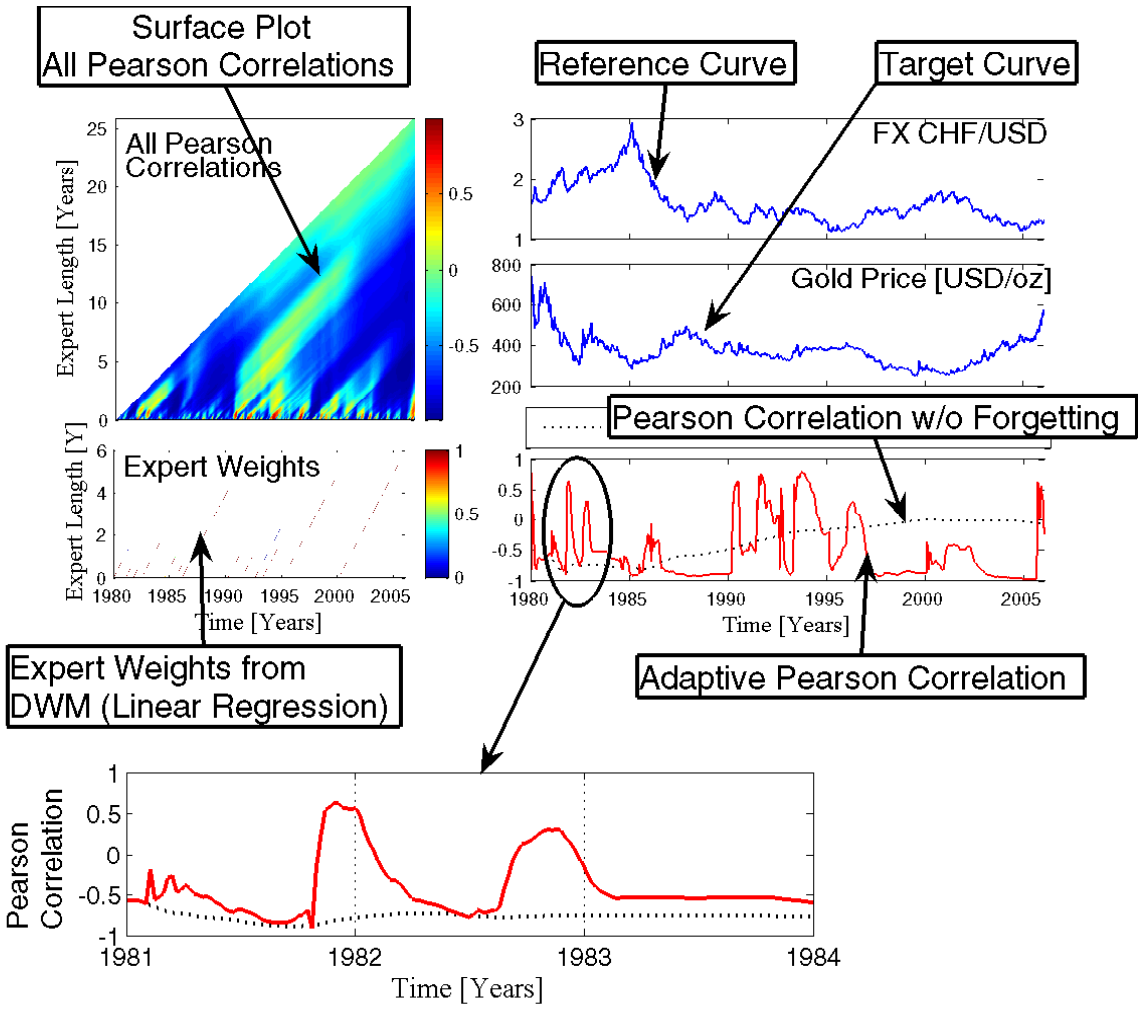


Figure 6.1: Interpretation instruction.

In Figure 6.1 the five sub figures are marked by text arrows. All of these sub figures are explained in detail below:

- **Reference Curve**
  The sub figure in the upper right corner shows the curve of the Foreign Exchange CHF/USD which is the reference for all of our calculations. Therefore, this sub plot stays the same for all of the following figure blocks.

- **Target Curve**.
  The target curve is the curve of interest and sketched below the reference curve. The abscissa spans the same time range as the reference curve (1980-2006). Sometimes not all data is available. Then, we plotted the missing values as zero-values, but the time-range remains the same.

- **Surface Plot**
  The surface plot in the left upper corner shows all possible Pearson correlations between the reference curve and the target curve for all time steps and for all possible experts' window size lengths. On the axis of abscissa the full time range is depicted and on the axis of ordinates the experts' window sizes. For example, the point at year 1995 with a length of 5 years corresponds to the Pearson correlation value calculated on a window of the range from 1990 until 1995. In the Figure the Pearson correlation value for this point is about 0.2 which can be derived by matching the point's color with the values on the color bar on the right.

  The topology in the correlation surface plot reveals the deeper relations between the two explored curves. So, we see how stable the correlations are, where changes occur, and where patterns re-emerge. In this example we see that a strong negative correlation dominates the most of the surface plot. The exceptions are the first 7.5 years of the nineties where the negative correlation switches sporadically to a positive correlation.

- **Expert Weights**
  The sub plot on the lower left corner illustrates the expert weights calculated by the dynamic-weighted-majority DWM ensemble algorithm. In the Sections 3.3.3 and 4.3.3 we discussed this kind of plots for the meteorology dataset. In short, the layout is the same as seen in the surface plot above, with the difference of having expert weights instead of Pearson correlations. The points of non-zero weights show the experts in power and which Pearson correlation values have to be considered from the surface plot above in order to calculate the adaptive Pearson correlation.

- **Pearson Correlation Curves**
  The sub plot in the lower right corner shows the Pearson correlation between the reference and the target curve.

The solid line is the final outcome of our work: the adaptive Pearson correlation. This curve has been calculated by selecting the most suitable (according to the expert weights) Pearson correlations from the surface plot. This curve is expected to reflect the correct correlation at each time step.

The dotted line shows the Pearson correlation based on all available data since the beginning of both curves. So, all past data is incorporated and we call this "Pearson correlation without forgetting". The purpose of this curve is to demonstrate the advantage of the adaptive over the non-adaptive Pearson correlation curve.

Keep in mind the time range is 26 years. So, narrow peaks in our graph correspond to larger periods. For example see the magnification in Figure 6.1 (the lowest plot). On the larger scale the two original narrow peaks turn out to be separated by about one year and are of a duration of almost a half year. So, this is considered as a real signal and not as an outlier. Of course, the finance experts get reports with higher resolution.

Pay attention to the fact that we are dealing with correlations which do not imply causality (see Section "Causality", p. 16)!

For the current calculations we used an update period for the DWM model of 5 days, i.e. about one week. This cycle turned out to be sufficient regarding the order of magnitude of the changes we are faced with. For even more fine-grained investigations shorter periods are feasible without limitations.

The illustration in Figure 6.1 provides valuable information for finance experts. But a static figure has it's limitations in the illustration of the dynamics of such a system. Dynamics are very important for the interpretation and intuitive comprehension since our problem is a *temporal* data mining problem subjected to fundamental time-dependent changes. Therefore, we provide an animated version of the illustration. The animated illustrations can be downloaded as movies under www.regimedrift.com/movies. The two screen shots in Figure 6.2 demonstrate how the animated illustration looks like. The upper plot is the screen shot of 1990 and the lower plot is the same, but one year later. The most eye-catching feature is the square covering the two variables of interest. The right border of the square is the actual time. The horizontal range is the illustration of the window size in power. As window size range illustration we have chosen to take the window size of the ensemble expert with the highest weight. Comparing the two screen shots reveals that the window size of the lower figure collapsed after one year. The adaptive Pearson correlation is drawn until the actual point of time. In the lower sub figures on the left a red circle shows the current ensemble expert with the highest weight. In the upper left sub figure a white circle shows the dominant Pearson correlation chosen by the ensemble expert with the highest weight.

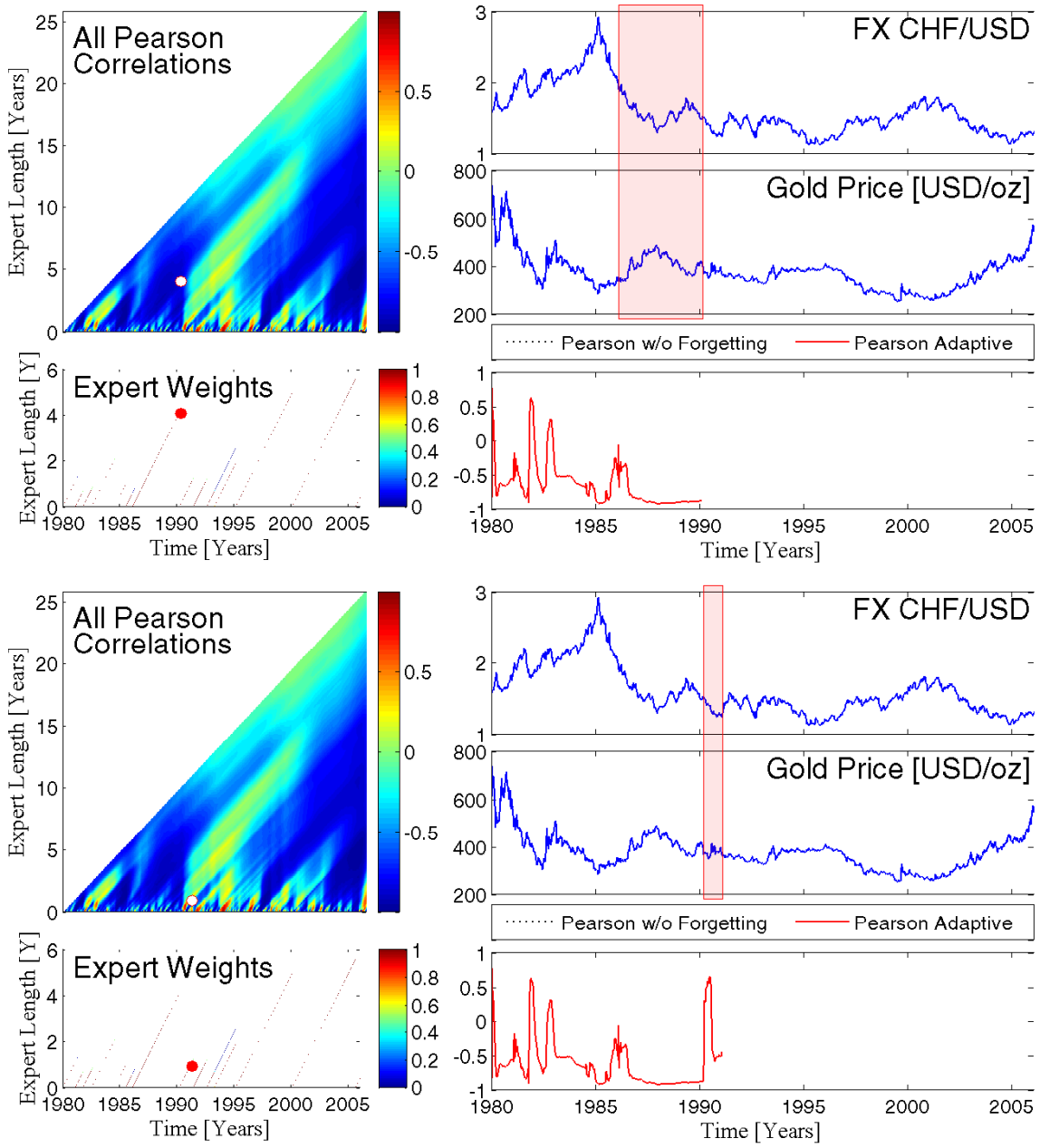


Figure 6.2: Two screen shots of the animated illustration of the regime calculation. The upper screen shot has been captured at 1990 and the second screen shot one year later at 1991.

6.3 Results

Now, it's time to apply our research on the real-world finance data. Together with the results we provide a short quote about the meaning of the variable. Each variable examination is conducted for the three regions of interest.

6.3.1 Foreign Exchange (spot)

As mentioned in the problem definition our main interest is the FX CHF/USD target. Our regions of interest are the regions with the currencies CHF (Swiss franc), USD (US dollar), and the EUR (European euro). Therefore, we have to look at the other exchange rates between these three regions, the FX EUR/USD and the FX CHF/EUR. Spot transactions are transactions in which currency is exchanged directly.

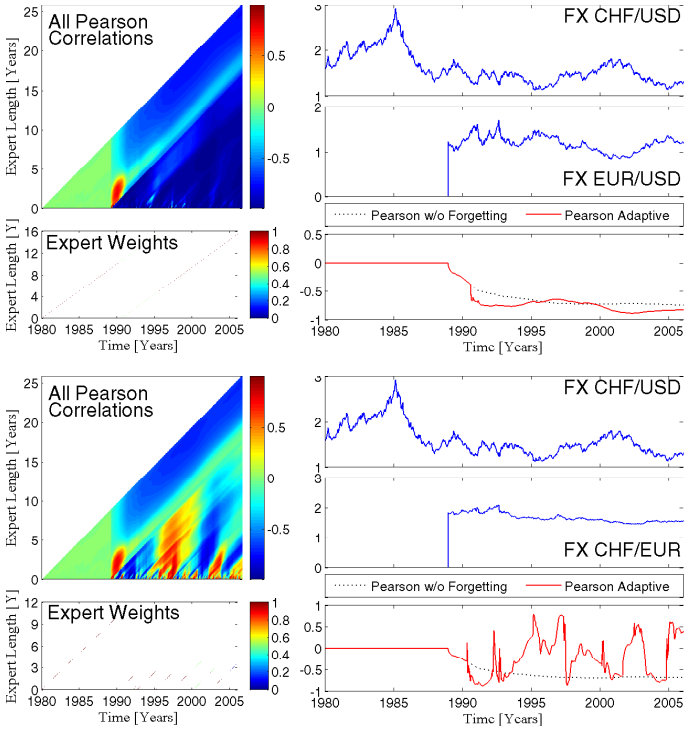


Figure 6.3: Foreign exchange (spot).

The adaptive Pearson correlation in Figure 6.3 shows a very high correlation between the FX CHF/USD and the FX EUR/USD. The reason for that behavior is the tight CHF/EUR relationship which can also be observed in the FX CHF/EUR raw data. The adaptive Pearson correlation between the FX CHF/USD and the FX CHF/EUR reflects the alternating relation between EUR and USD.

6.3.2 Forward Foreign Exchange Rate

Transactions consisting of contracts to exchange one currency to another at a future date, but terminated now are called forward transactions. Their exchange rate is called forward exchange rate.

As Figure 6.4 shows the forward rates are very similar to the foreign exchange rates and so are the adaptive Pearson correlations.

This effect is caused by the following situation. At the time of contract signing the positions have to be covered. *"Dealers in forward exchange usually balance their commitments; for instance, a contract to deliver forward marks can be offset against one to deliver forward dollars, and nothing more has to be done about it. If a particular dealer cannot manage this he will be in communication with another who may be in the opposite position. It may not, however, always be possible to offset every transaction. If this is not done, the dealer must make a spot purchase of the currency in excess demand in the forward market. If he did not do this he would risk an exchange loss on some of his forward transactions".* [Encyclopædia Britannica, 2007]. Thus, the forward exchange rates reflect the current (spot) exchange rates.

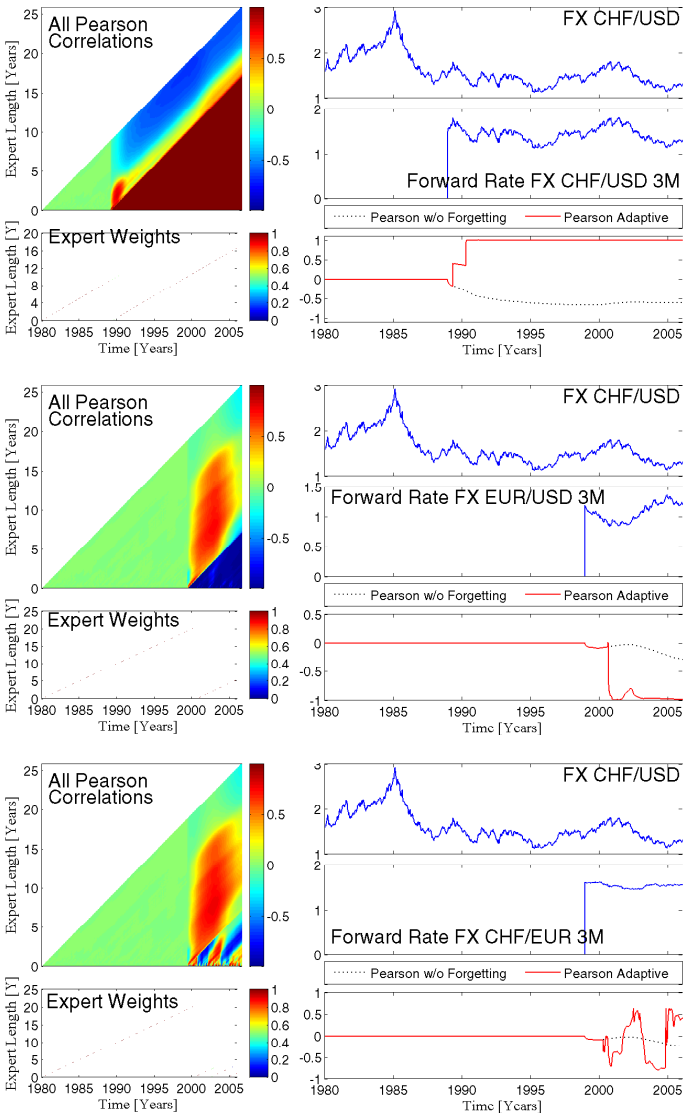


Figure 6.4: Forward foreign exchange rate.

6.3.3 Currency Swap

A swap is in general *a financial term for a combined or simultaneous buying and selling operation [UBS Dictionary of Banking, 2007].*

1. *Swaps between central banks: transactions that are frequently carried out in connection with the International Monetary Fund IMF or the Bank for International Settlements BIS to bridge international liquidity crises.*

2. *Capital-market swaps: agreements whereby the two parties undertake to swap payments over a specified period on specified dates and at conditions fixed in advance. The swap contract can either refer to the exchange of interest payments (interest-rate swap) or the exchange of interest payments and nominal amounts in different currencies (currency swaps).*

3. *Synonym for currency swap.*

4. *Synonym for debt-equity swaps.*

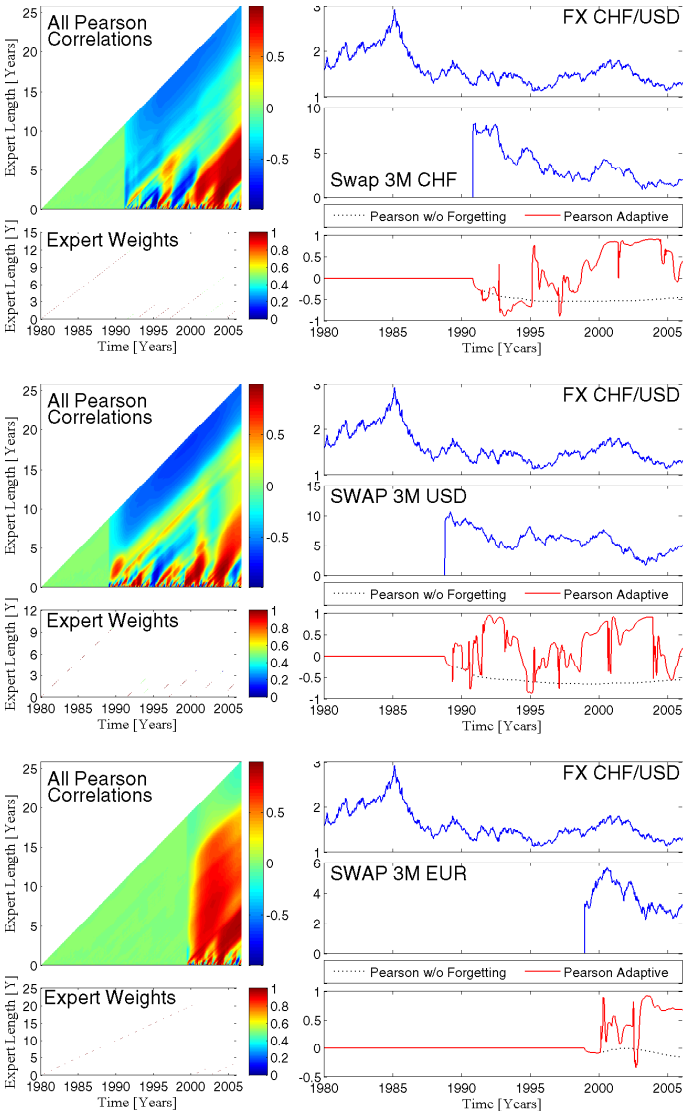


Figure 6.5: Currency swap.

Here we are dealing with currency swaps (bullet item 2) which are according to the Bloomberg glossary [Bloomberg Financial Glossary, 2000] *an agreement to swap a series of specified payment obligations denominated in one currency for a series of specified payment obligations denominated in a different currency.*

The curves in Figure 6.5 show the 3M (three month) swaps for CHF, USD, and EUR. Noticeable is the increase of the correlation in the late nineties up to a high correlation between the swaps and the exchange rate CHF/USD.

6.3.4 Commodities

We limit our study on commodities to the gold, silver and oil price. All three resources are accounted in USD. The oil price is the price of Brent Crude which is sourced from the North Sea.

Whereas silver has not a pronounced correlation to the exchange rate CHF/USD, the gold and oil price have a high correlation most of the time. The correlations to gold and oil are switch from strong positive to strong negative. Looking at the decade starting in the early nineties, gold and oil price are complementary to each other with respect to the reference FX CHF/USD. Then after 2003 both commodities have again a high negative correlation to the FX CHF/USD.

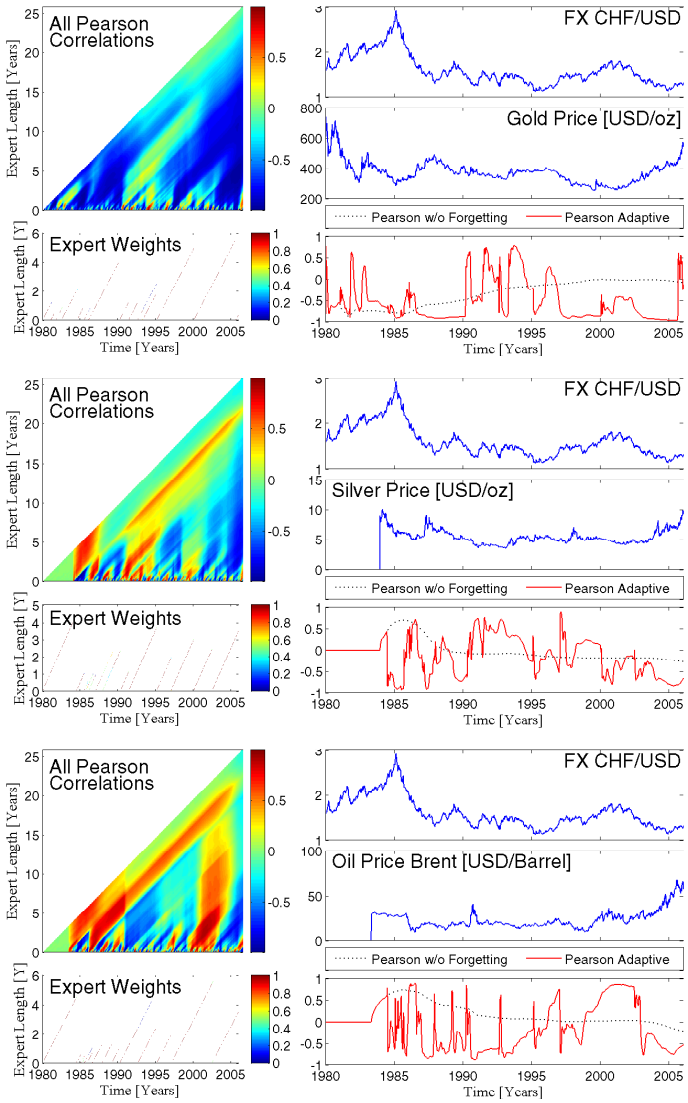


Figure 6.6: Commodities: gold, silver, and oil.

6.3.5 Interest Rates

Key Interest Rates

The key interest rate is also known as key rate, base rate (GB), and prime rate (USA). The key interest rate is set by a central bank for central bank funds. The economic importance of the key interest rate is its fundamental role as monetary policy instrument. This is described by the SNB: "*All regular monetary policy instruments of the SNB are based on repo transactions. In a repo transaction, the cash taker sells securities spot to the cash provider. At the same time the cash taker enters into an agreement to repurchase securities of the same type and amount from the cash provider at a later point in time. The cash taker pays interest (the repo rate) for the duration of the transaction. From an economic perspective, a repo is a secured loan. Regular instruments are divided into main financing operations and liquidity absorbing operations, fine-tuning operations, as well as the intra day facility and the liquidity-shortage financing facility.*"[SNB Glossary, 2007].

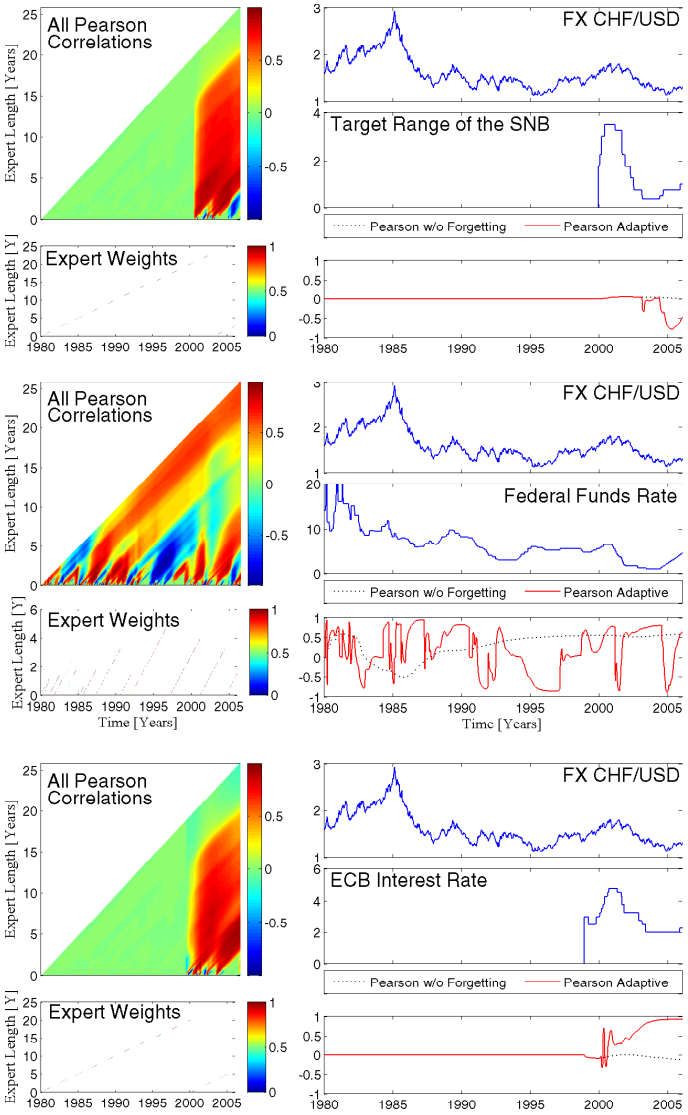


Figure 6.7: Key interest rate.

Due to the direct correlation of the key interest rate to the monetary liquidity the correlation to the exchange rates in Figure 6.7 is high too (except for the time range where no data has been available).

LIBOR

"The London Interbank Offered Rate (LIBOR) designates the interest rates fixed every business day at 11:00 a.m. (London time) by the British Bankers' Association. These are the rates at which major banks are prepared to grant unsecured money market loans to each other. The LIBOR is fixed according to a clearly defined procedure for different currencies and maturities. The Swiss franc LIBOR corresponds to the average of the current interest rates of six leading banks." [SNB Glossary, 2007].

The LIBOR development in Figure 6.8 is parallel to the development of the key interest rates discussed above. Therefore, the adaptive Pearson correlation curves look the same as above.

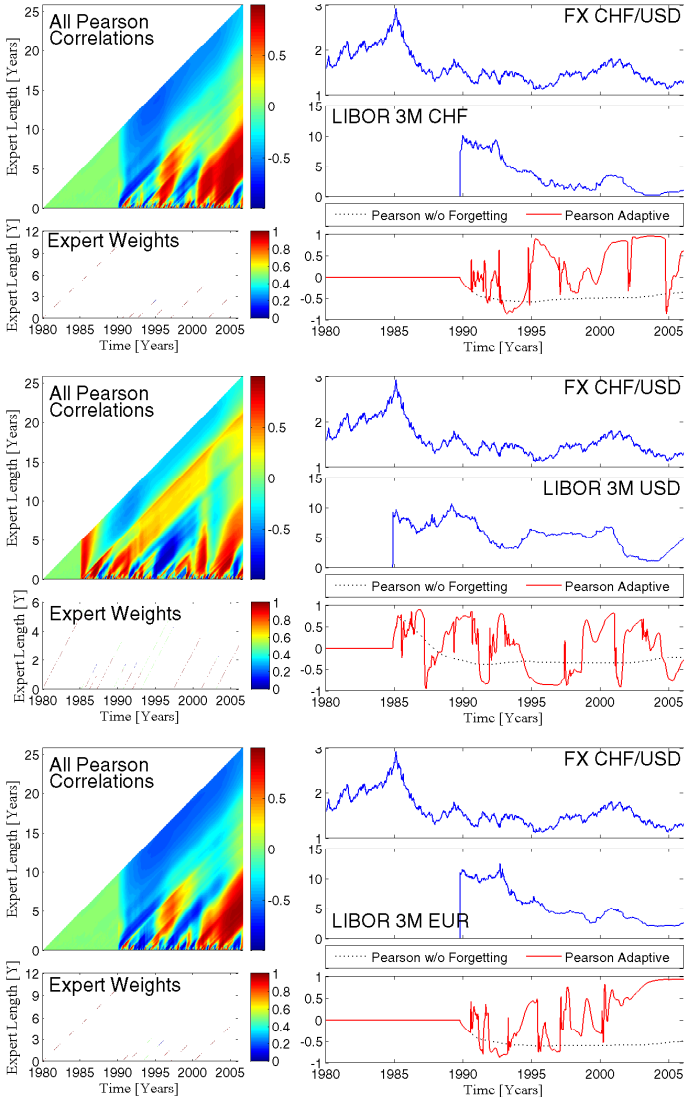


Figure 6.8: LIBOR.

Treasury Bonds

Figure 6.9 shows the Federal bond (CH), Treasury securities (US), and the euro-denominated bond curves, where the Treasury bond of Germany has been taken as representative for the EU. For all these bonds the statement of the SNB is true: *"A Federal bond is a fixed-interest debt certificate (bond issue) of the Swiss Confederation employed by the Confederation for medium-and long-term borrowing in the capital market."* [SNB Glossary, 2007]

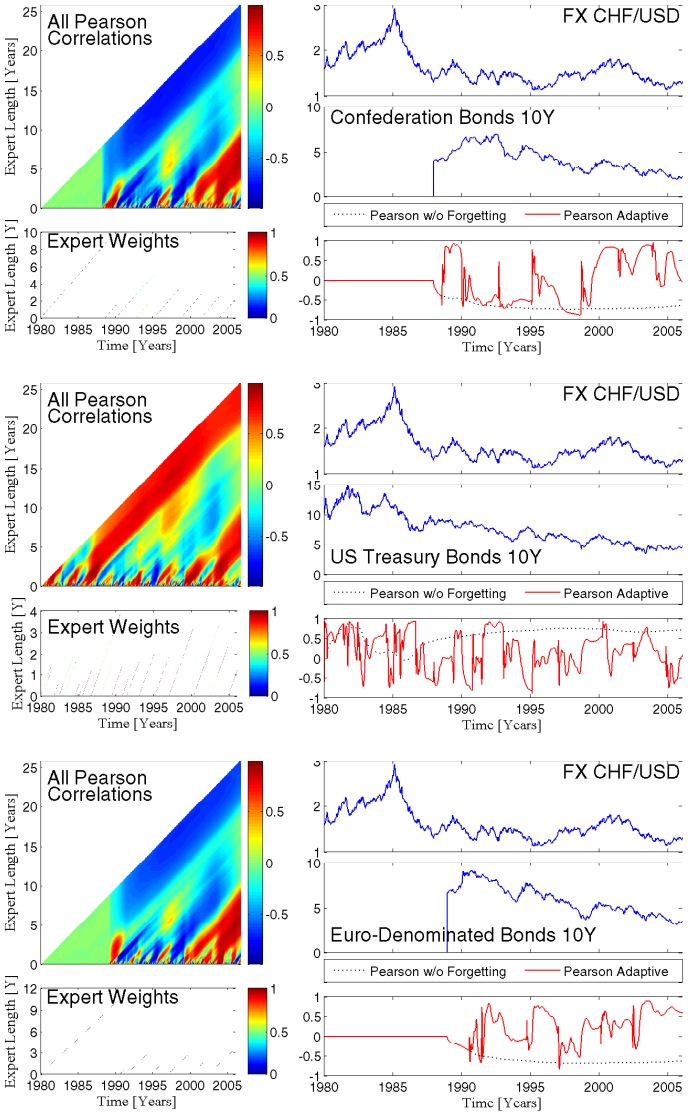


Figure 6.9: Treasury bonds.

The UBS Dictionary of Banking provides more information about Treasury securities. *"U.S. Treasury securities are debt obligations of the U.S. government and, as such, are backed by the "full faith and credit" of the U.S. government. Considered the safest of all investments, they are viewed as having virtually no credit risk. As a result of this safety, treasuries generally offer the lowest rates of all widely traded debt in the domestic market. The U.S. Treasury market is the most liquid debt market in the world, offering the most efficient trading and pricing. Treasuries are exempt from state and local taxes and are issued as:*

- *Bills: Issued in maturities of no more than 6 months. Sold at discounts to their value at maturity (i.e., par amount).*
- *Notes: Typically issued in 2, 3, 5 and 10 year maturities. Interest paid semi-annually.*
- *Bonds: Issued in maturities from 10 to 30 years and interest is paid semi-annually.*
- *Zeros: Represent ownership of a future interest payment on a Treasury note or bond..."*

The maturities for our three variables under observation are 10 years. The bond curves of all three regions do not differentiate much at first sight. But looking at the Pearson surface plots and the resulting adaptive Pearson correlation reveals some differences – specially, between the US and the two European regions. The adaptive Pearson correlation is very spiky for the US treasure securities. The correlation for the Federal and the euro-denominated bonds are smoother.

The second bond futures (Fig. 6.10) are plotted in units "$100 - value$". Therefore, they look mirrored to the corresponding Treasury bonds (Fig. 6.9). Except for the mirroring, the second bond future curves are parallel to the Treasury bond curves and so are the adaptive Pearson correlations.

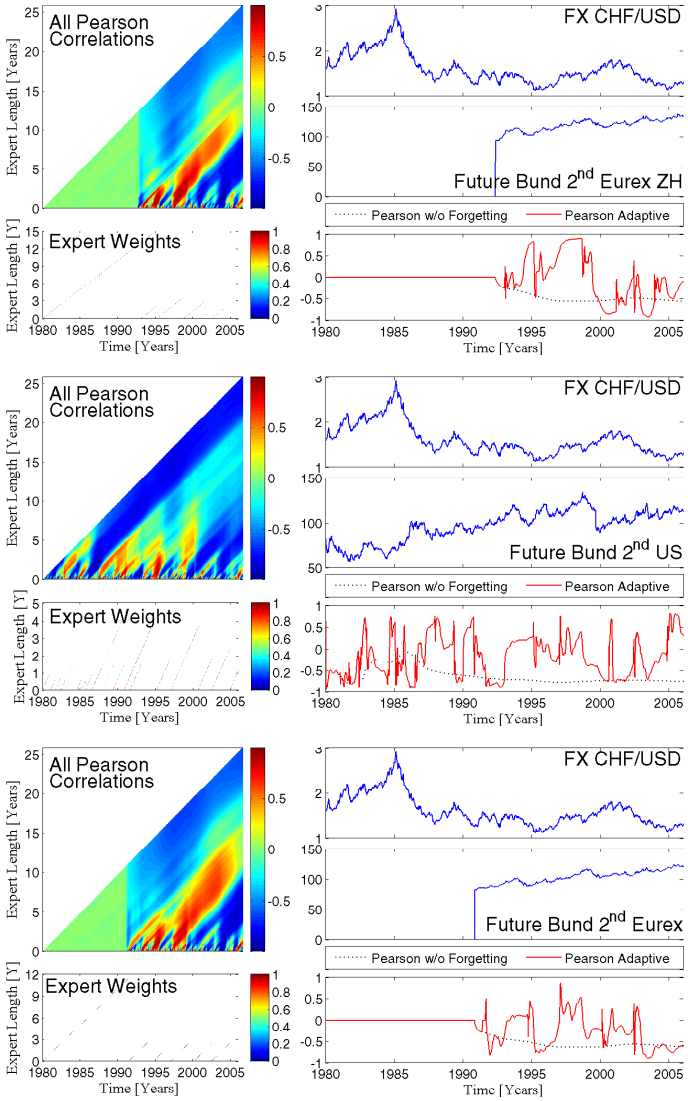


Figure 6.10: Future bonds 2^{nd}.

Forward Rate Agreements

[UBS Dictionary of Banking, 2007] defines the forward rate agreement (FRA) as follows. *A FRA is a forward transaction between banks and industrial firms whereby the two parties agree on an interest rate for a future period and no initial margin payment is required at the time the contract is concluded. Unlike financial futures, FRAs are not standardized and are not traded on interbank markets. They are used to hedge the risk of a change in interest rates by locking in the current interest rate for future payments.*

In the first column of Figure 6.11 shows the 3x6 forward rate agreements and the second column shows the 3x9 FRAs for the three currencies CHF, USD, and EUR. The designation 3x9 stands for hedging of the interest rate under a contract that begins in three month's time and remains in force for another six months (nine months after issue).

The adaptive Pearson correlations show that there is no significant difference between the 3x6 and the 3x9 FRAs. The CHF FRAs show a higher correlation to the exchange rate CHF/USD than the USD and EUR FRAs. The peak in the USD 3x9 FRA is an outlier. If eliminated, the curve looks the same as the USD 3x6 FRA curve.

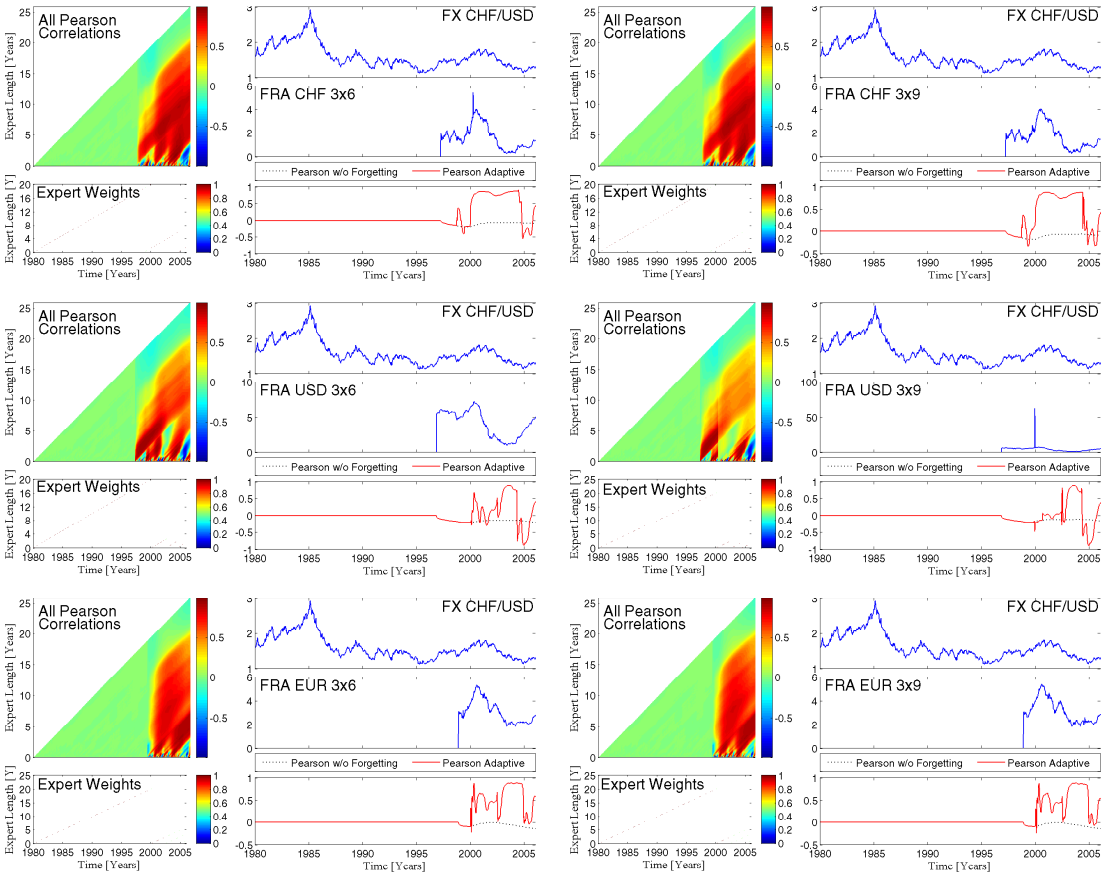


Figure 6.11: Forward rate agreements.

As in real-world applications we use unprocessed real-world data for our survey. Nevertheless, our algorithm demonstrates its robustness towards such disturbing factors on the USD 3x9 FRA variable.

Futures Short-Term Interest

Figure 6.12 shows the futures of the short-term interests. The short-term interests are issued by banks. The short-term interest description "3M 1^{st} Generic" for example is defined as follows. "3M" stands for a maturity of 3 months and "generic" stands for the continuous illustration by taking the next future contract after expiration of the preceding future. "1^{st}" stands for the futures expiring as next and "2^{nd}" for the futures expiring after the next one, thus, more distant.

There is almost no difference between the "1^{st}" and the "2^{nd}" future correlation curves, except a slight difference between "1^{st}" and the "2^{nd}" US futures. There are more pronounced differences in the adaptive Pearson correlation between the different regions.

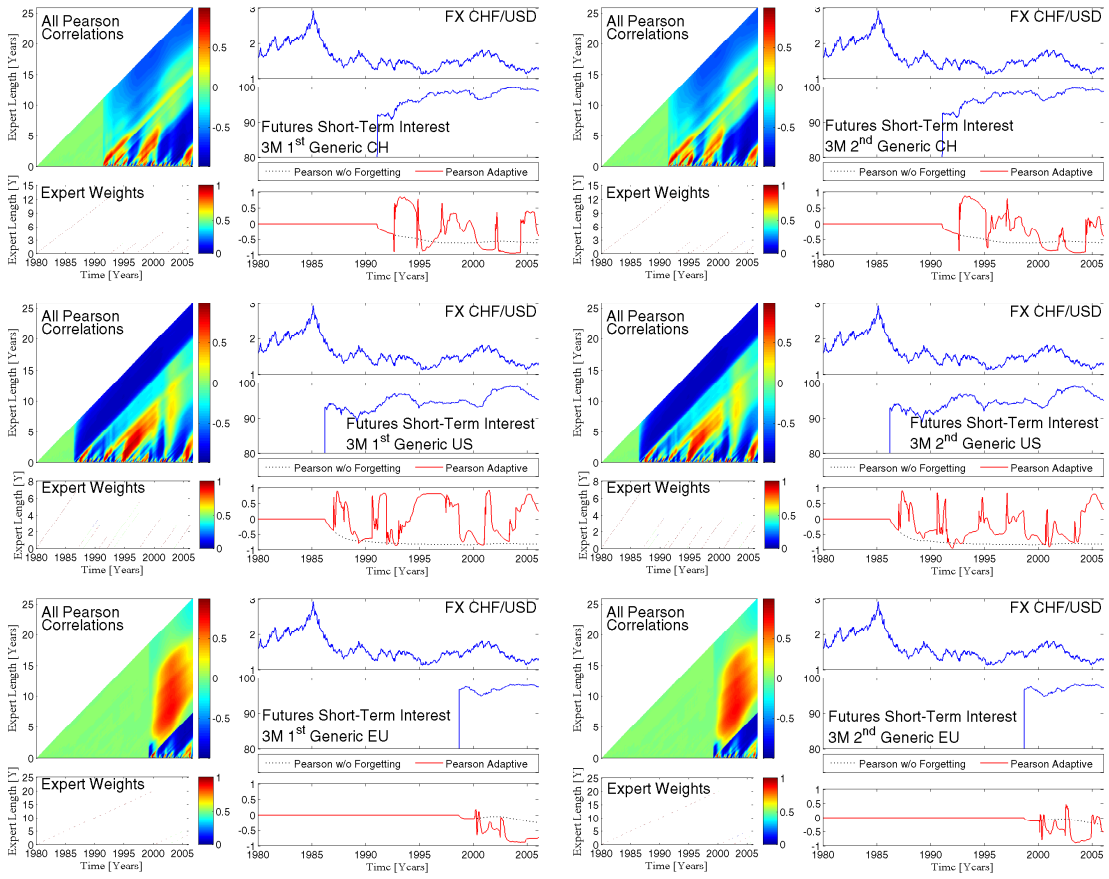


Figure 6.12: Futures short-term interest.

6.3.6 Stock Exchange

A stock exchange index or share index is an indicator showing changes in the average prices of shares or groups of shares on the stock market. We have chosen a representative stock exchange index for each of the three regions CH, US, and EU. In Switzerland the most commonly used equity price indices are the Swiss Performance Index (SPI) and the Swiss Market Index (SMI). The UBS Dictionary of Banking states the SPI is the "*broadest based Swiss share index, covering all domestic companies listed on the Swiss Exchange. Weighted by capitalization and dividend-adjusted, the SPI is an ideal benchmark for performance comparisons.*" So, we decided to go with the SPI. For the US region we preferred the NASDAQ to the Dow Jones because of its composition towards more innovative stocks. NASDAQ is the "*acronym for National Association of Securities Dealers Automated Quotations. US electronic exchange for high-growth, innovative stocks, catering for OTC traders.*" The German DAX has been chosen as representative for the EU. The DAX (Deutscher Aktienindex) is a "*stock index, which measures the performance of the 30 largest German companies in terms of order book turnover and market capitalization.*" [UBS Dictionary of Banking, 2007]

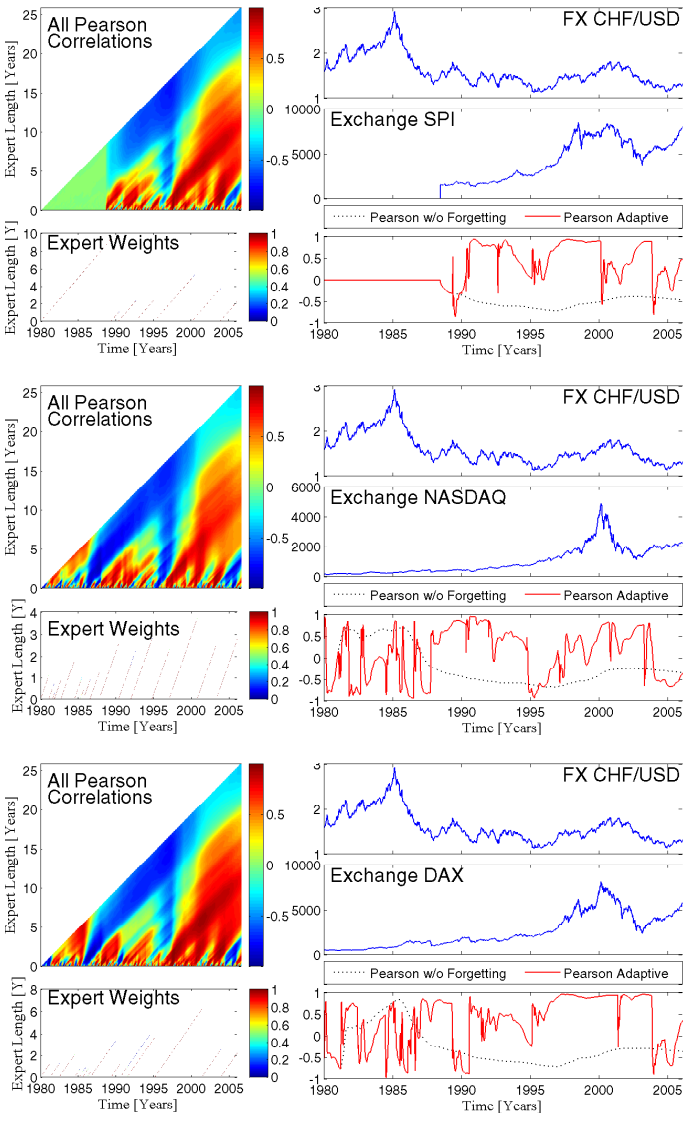


Figure 6.13: Stock exchange index

All stock indices have a climax in 2000 due to the dot-com bubble (Fig. 6.13). Even though, there is an overall difference in the shape between the share indices and the exchange rate CHF/USD, we observe a high correlation between these curves. After the late eighties we observe higher and more stable correlations.

Now, we take a look at equities futures in Figure 6.14. As representative for the stock futures in the regions of interest we have chosen the futures in three trading places of these regions. 2^{nd} stands for the futures that will be converted after the first future conversion date. The raw curves are similar to the stock exchange index except for the US region, where the dot-com bubble peak is not as pronounced. The reason for this effect is that the future curve is based on the Dow Jones which not that sensitive to new technology valuation changes NASDAQ. Comparing the result with the results of the stock exchange indices, the adaptive Pearson correlation is about the same for the CH and EU and also for the US region (when neglecting the constant value region in the nineties).

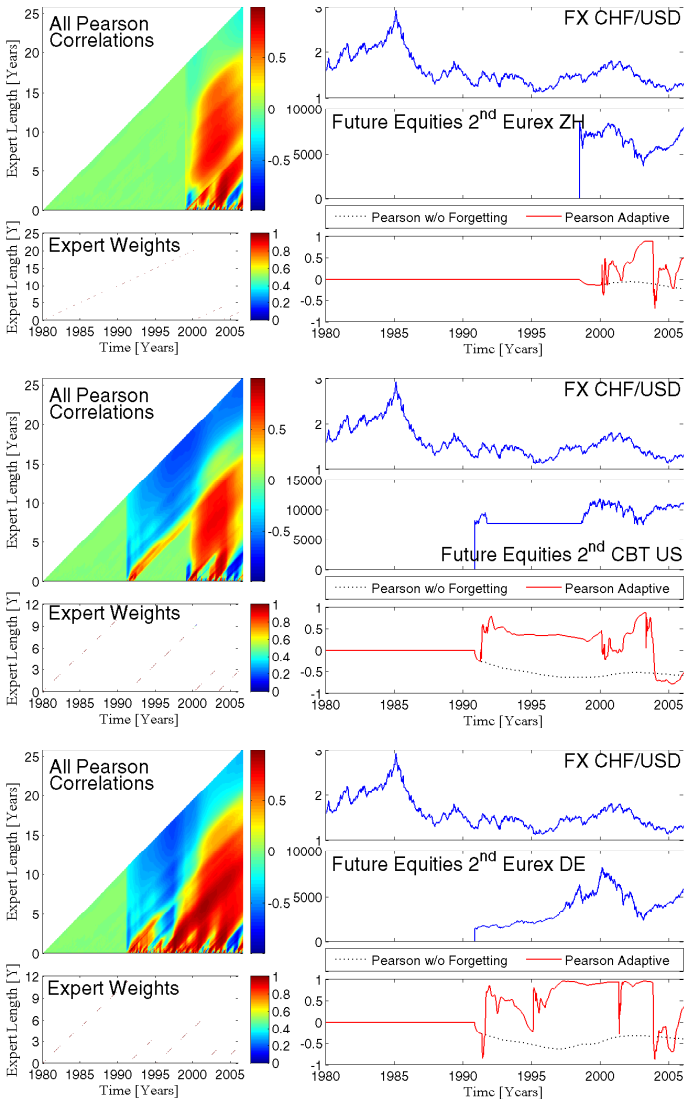


Figure 6.14: Future equities 2^{nd}.

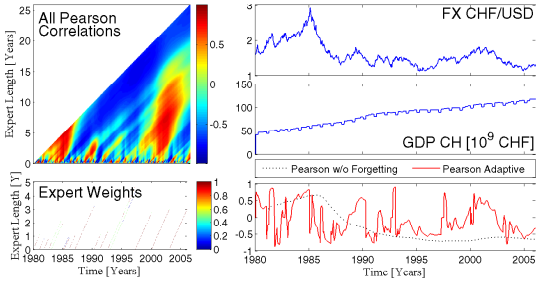
6.3.7 Gross Domestic Product

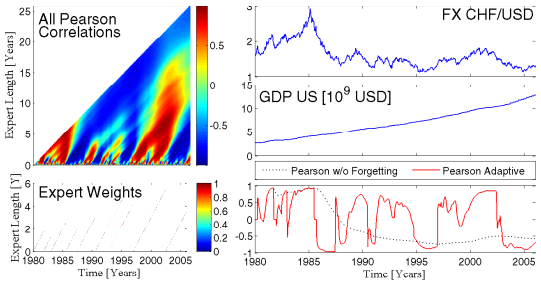
The Gross Domestic Product (GDP) is defined as the "*total market value of the goods and services produced by a nation's economy during a specific period of time. It includes all final goods and services - that is, those that are produced by the economic resources located in that nation regardless of their ownership and that are not resold in any form*." [Encyclopædia Britannica, 2007]. We distinguish between two kinds of the GDP. The nominal and the real GDP. The nominal GDP is calculated using the actual price level and is affected by inflation. The real GDP is inflation-adjusted. Figure 6.15 shows the calculations for the nominal GDP and Figure 6.16 for the real GDP, respectively.

GDP values are published quarterly. Therefore, the values appear cascaded (in particular the CH real GDP values which are subjected to an annual inflation). Bloomberg classifies the GDP as highly market important.

As Figure 6.15 and 6.16 show, all three economic regions Switzerland, US, and EU have a continuous GDP growth. Typically, the values of the European Union are only available since the nineties.

Looking at the surface plots on the left side of each of the Figure units shows that all Pearson correlation plots are very similar for the adjusted and non-adjusted case. Thus, the adaptive Pearson correlation curves look very similar, too. The Pearson correlations show switching regimes from high negative correlations to high positive correlations.





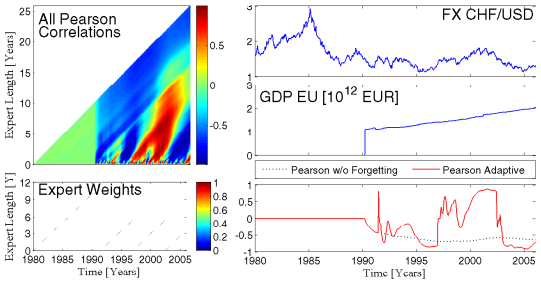
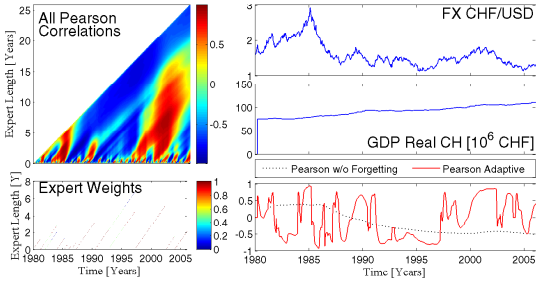
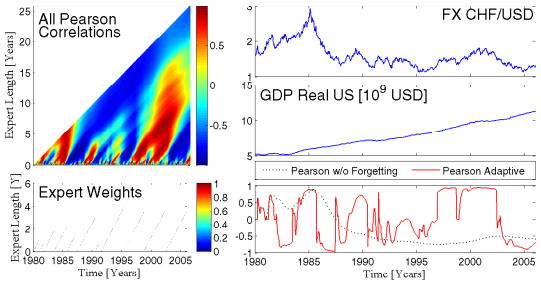


Figure 6.15: Gross Domestic Product, nominal.





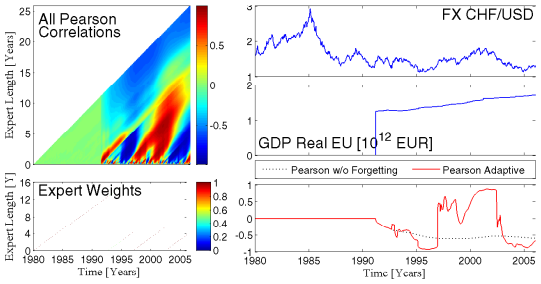


Figure 6.16: Gross Domestic Product, real.

6.3.8 Money Supply

[Encyclopædia Britannica, 2007] defines money supply as "*the liquid assets held by individuals and banks. The money supply includes coin, currency, and demand deposits (checking accounts). ... The Federal Reserve Board in the United States and the Bank of England in the United Kingdom regulate the money supply to stabilize their respective economies. The Federal Reserve Board, for example, can buy or sell government securities, thereby expanding or contracting the money supply.*"

The Swiss National Bank SNB differentiates between four kinds of money supplies M0, M1, M2, and M3 [SNB Glossary, 2007].

- M0 is the money supply of the central bank, also referred to as the monetary base, or occasionally as the cash base.
- M1 comprises currency in circulation in the form of Swiss francs (banknotes and coins) held by the public plus sight deposits in Swiss francs held by the resident public at banks and the post office as well as transaction deposits.
- M2 is defined as the sum of the money stock M1 and savings deposits. Excluded from savings deposits are pension fund monies invested in schemes with restricted terms and tax benefits within the framework of the mandatory occupational pension scheme (pillar 2) and the voluntary, individual pension scheme (pillar 3).
- M3 comprises the money stock M2 plus time deposits.

Bloomberg classifies the market relevance of this kind of data as low.

For our research we have chosen to focus on M1, M2, and M3. The columns in Figure 6.17 show the calculations for different kinds of money supply for each economic region (rows).

First, we have a look at the raw money supply curves. The behavior of the central banks of the different economic regions seems to be a little bit different. Whereas the most liquid kind of money increases continuously in the EU, the short-term available money in the US and CH changes sporadically over time. The overall amount of money M3 is continuously increasing for all of the three economic regions.

Second, we look at the adaptive Pearson correlations. Comparing the curves for Switzerland M1 and M2 the interval between 1992 and 2000 is eye-catching. The correlation between the foreign exchange rate CHF/USD for M1 is much more pronounced than for M2. This behavior is caused by the money bound by the pensions in M2, so that investors are not able to react as fast on exchange rate variations. Also the liquid M1 from the US correlates with the exchange rate because of the Dollar.

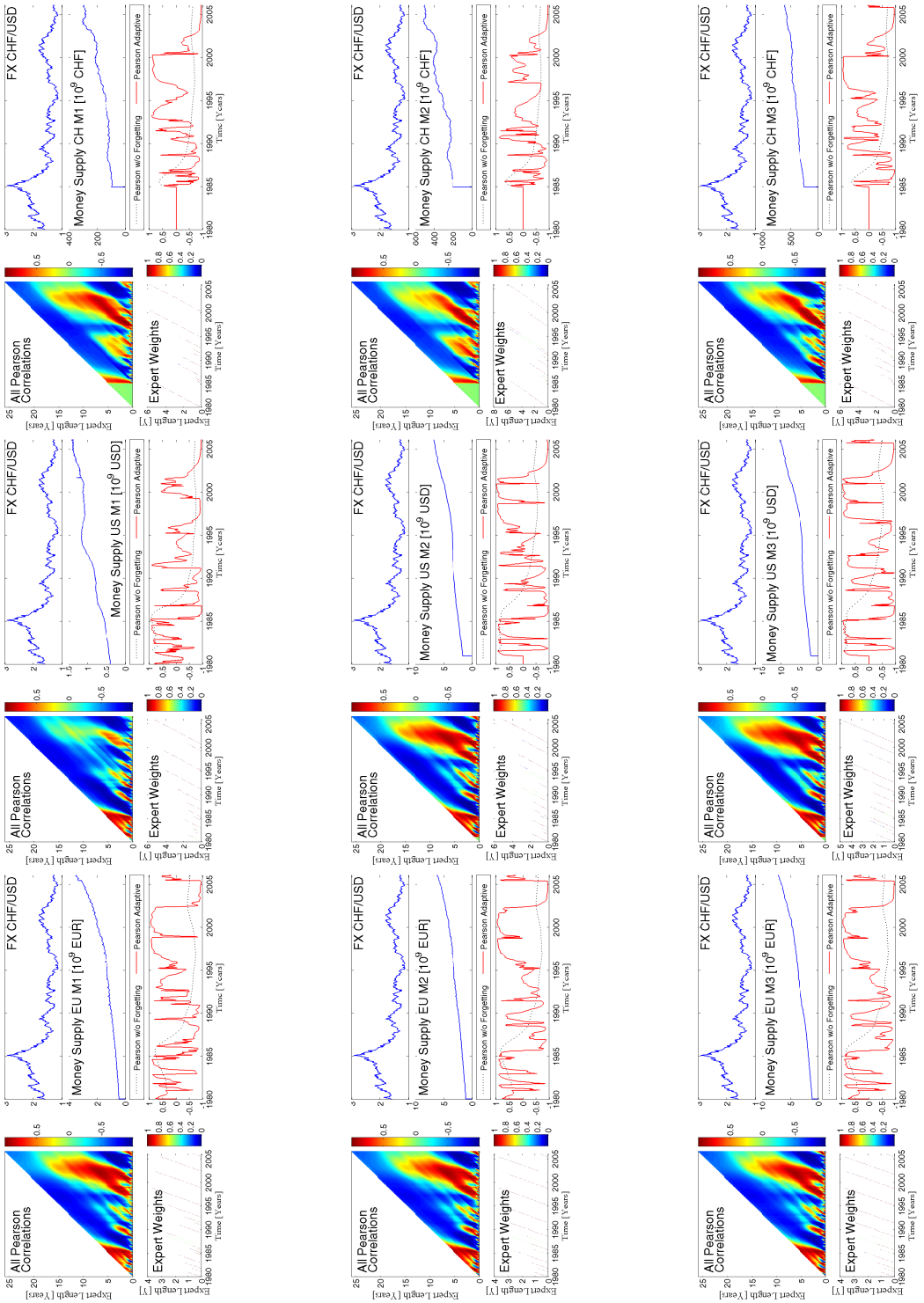


Figure 6.17: Money supply.

6.3.9 Consumer Price Index and Producer Price Index

[Encyclopædia Britannica, 2007] defines the Consumer Price Index (CPI) "*as measure of living costs based on changes in retail prices. Such indexes are generally based on a survey of a sample of the population in question to determine which goods and services compose the typical market basket. These goods and services are then priced periodically, and their prices are combined in proportion to the relative importance of the goods. This set of prices is compared with the initial set of prices (collected in the base year) to determine the percentage increase or decrease. Consumer price indexes are widely used to measure changes in the cost of maintaining a given standard of living.*" Bloomberg classification: very high importance.

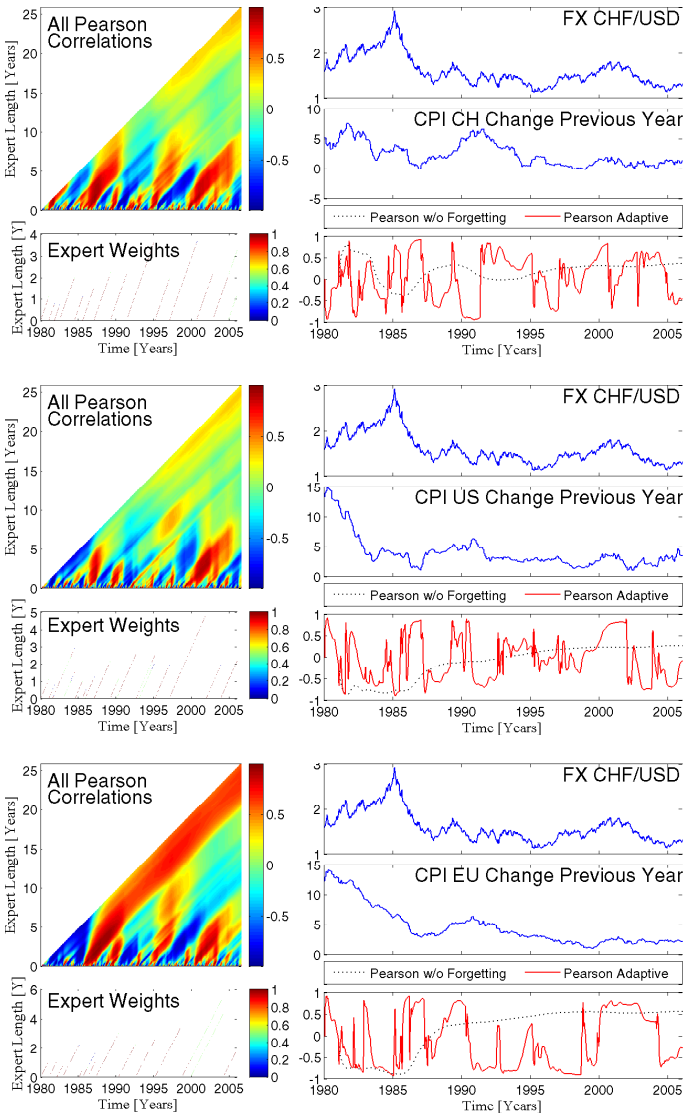


Figure 6.18: Consumer Price Index.

The U.S. Department of Labor defines in the "Bureau of Labor Statistics Handbook of Methods" the Producer Price Index (PPI). "*The PPI measures average changes in prices received by domestic producers for their output. Most of the information used in calculating producer price indexes is obtained through the systematic sampling of virtually every industry...*" [BLS Handbook of Methods, 2003].

Nouriel Roubini[2] provides some more information. "*The Producer Price Index (PPI) is the first indicator of inflation each month. It is a measure of wholesale prices at the producer level for consumer goods*

[2]Professor of Economics and International Business at the Stern School of Business, New York University, http://pages.stern.nyu.edu/ nroubini/bci/ProducerPriceIndex.htm (October 12, 2007)

and capital equipment. Unlike the CPI, it does not include services. It compares prices for approximately 3,450 commodities to a base period. Currently, the base period, which equals 100, is the average prices that existed in 1982." Bloomberg classification: high importance.

Whereas the CPI measures price changes from the consumer's perspective, the PPI measures it from the sellers and manufacturer's perspective. For more insight on the difference between the Producer Price Index and the Consumer Price Index we recommend the article [BLS, 2004] published by the Bureau of Labor Statistics.

The CPI curves in Figure 6.18 show high correlations for the CH and EU region. The Swiss region differs by an increase of the CPI in the early nineties. In this time range the correlation for the US region is very weak.

The PPI curves in Figure 6.19 show a similar behavior for the US and EU region. In the eighties the Swiss PPI correlation is very high and similar or even more robust as the other two regions. But in the early nineties the correlation of the Swiss region differs from the other regions. After that, the Swiss PPI correlation is not as pronounced as the PPI of the other two regions.

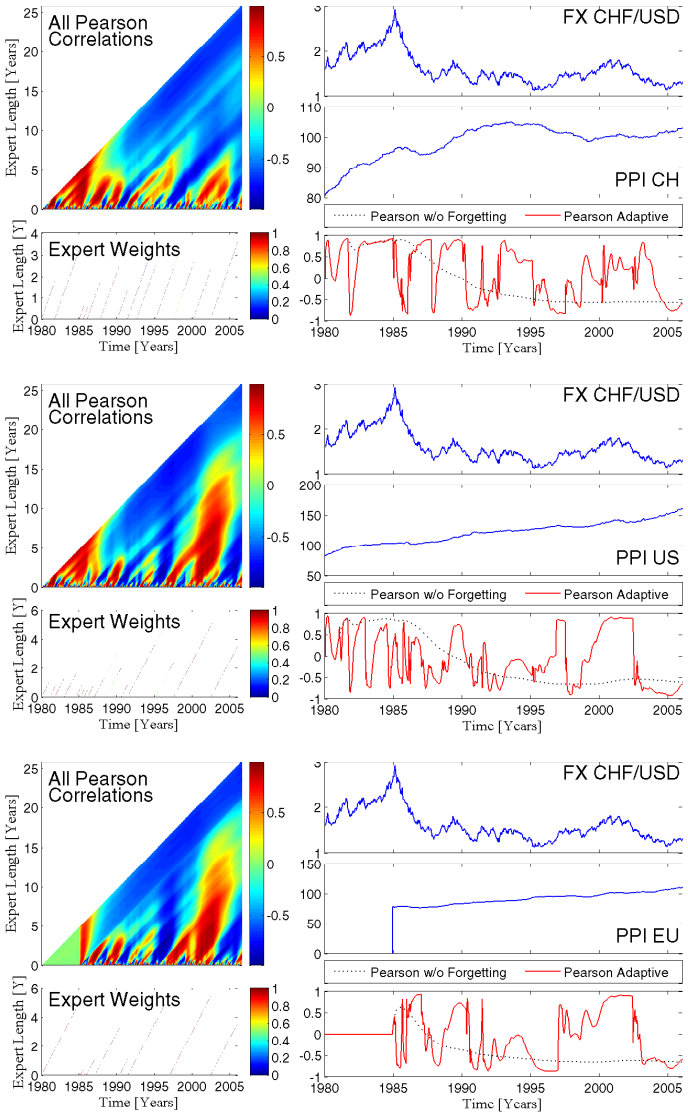


Figure 6.19: Producer Price Index.

6.3.10 Industrial Production Index

The industrial production index measures real production output of a region. We used the overall index whereas it can be obtained by market and industry groups. The data is published by the Federal reserve (US), Eurostat (EU), and the Swiss Federal Statistical Office (CH). The US and EU indices are published monthly and the CH index quarterly. The US reference value of 100% is based on the production output at the end of 2002. The index covers output, capacity, and capacity utilization in the U.S. industrial sector, which is defined by the Federal Reserve to comprise manufacturing, mining, and electric and gas utilities.

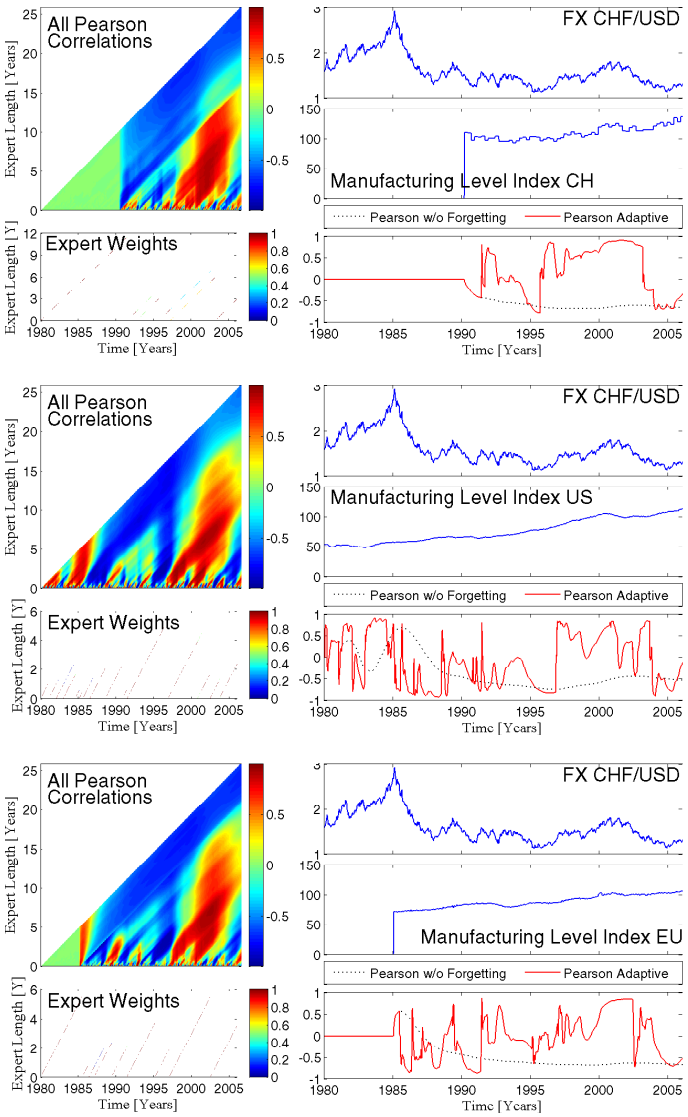


Figure 6.20: Industrial Production Index.

The adaptive Pearson correlation (Fig. 6.20) between the US and CH industrial production index and the CHF/USD exchange rate shows very high correlation after the year 1995. The correlation for the EU region is high after the year 2000. This shows the industry's capability to adapt its production output to the economic situation. In our opinion, this has been enabled by the support of information systems which allow more efficient supply chain management and just-in-time production.

6.3.11 Purchasing Managers Index

The Purchasing Managers Index PMI is defined by [Kennon, 2007]. "*The Purchasing Managers Index is released on the first day of the month by the National Association of Purchasing Managers. The PMI measures five factors in business: new orders, inventory levels, production, supplier delivers, and employment conditions. Each of these five factors are adjusted and weighed according to time of year and other events. A PMI over 50% means that manufacturing is growing and expanding. A PMI under 50% means that manufacturing is declining. A PMI of 42.7% or more over a long period of time means the economy as a whole is expanding. A PMI of 42.7% of below over a long period of time means the economy as a whole is contracting.*" Bloomberg classification: high.

The PMI for all three regions is quite the same and so is the adaptive Pearson correlation, see Figure 6.21. Overall the correlations are not as pronounced as maybe expected.

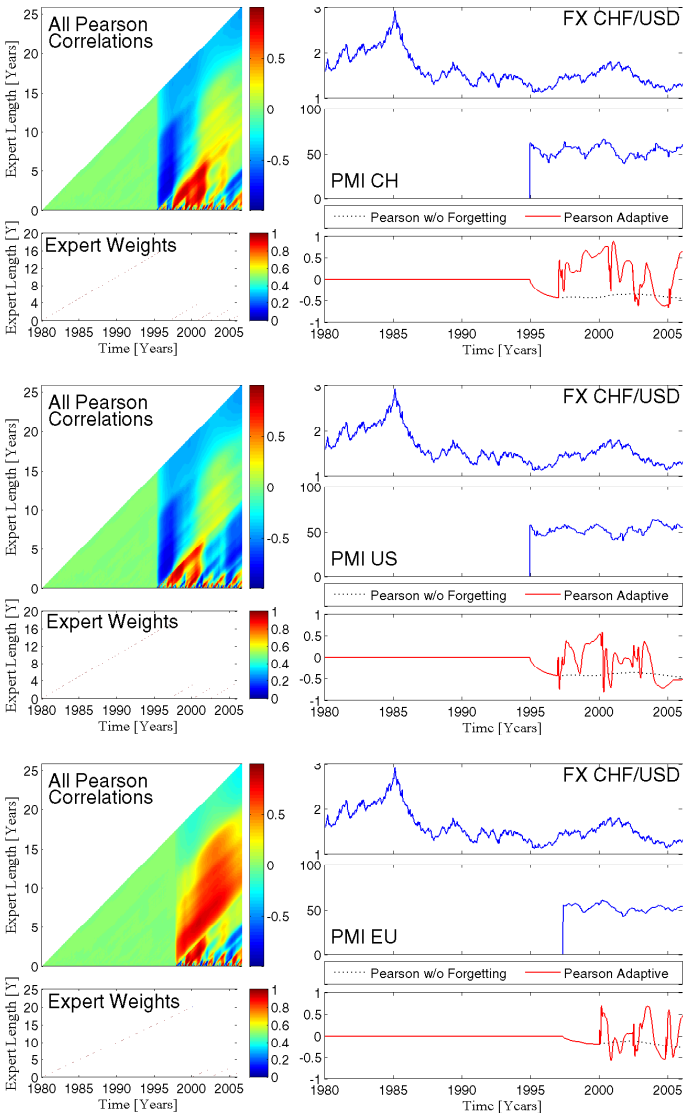


Figure 6.21: Purchasing Managers Index.

6.3.12 Unemployment Rate

Statistics on unemployment are collected and analyzed by government labor offices in most countries. Bloomberg classifies the employment rate as highly important for the market.

The raw variables in Figure 6.22 are plotted in percentage units with respect to the entire population. The unemployment rate differs for the three economic regions. The adaptive Pearson correlation between the exchange rate CHF/USD and the unemployment in CH is more stable than the correlation for the US unemployment rate. The EU region unemployment rate is similar to the Swiss rate, but not that pronounced. In our opinion, the high correlation between the Swiss unemployment rate and the exchange rate CHF/USD reflects the flexible employment market in Switzerland which allows to adapt to the current economic situation.

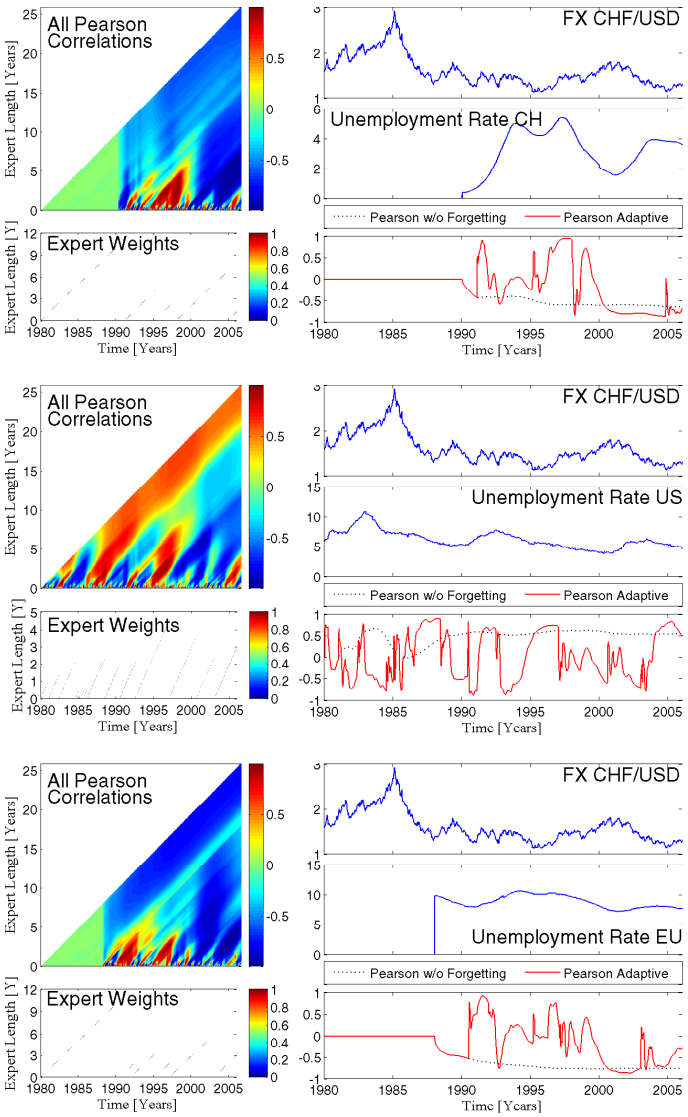


Figure 6.22: Unemployment rate.

6.3.13 Wages

We differentiate between nominal and real (inflation-adjusted) wages. We focus on real wages, since the higher the real wages the higher the purchasing power. Purchasing power has influence on consumption and savings behavior and thus, the economic situation. As representative for the EU wages we have chosen to use the UK wages. Bloomberg classifies the wages as medium important for the market.

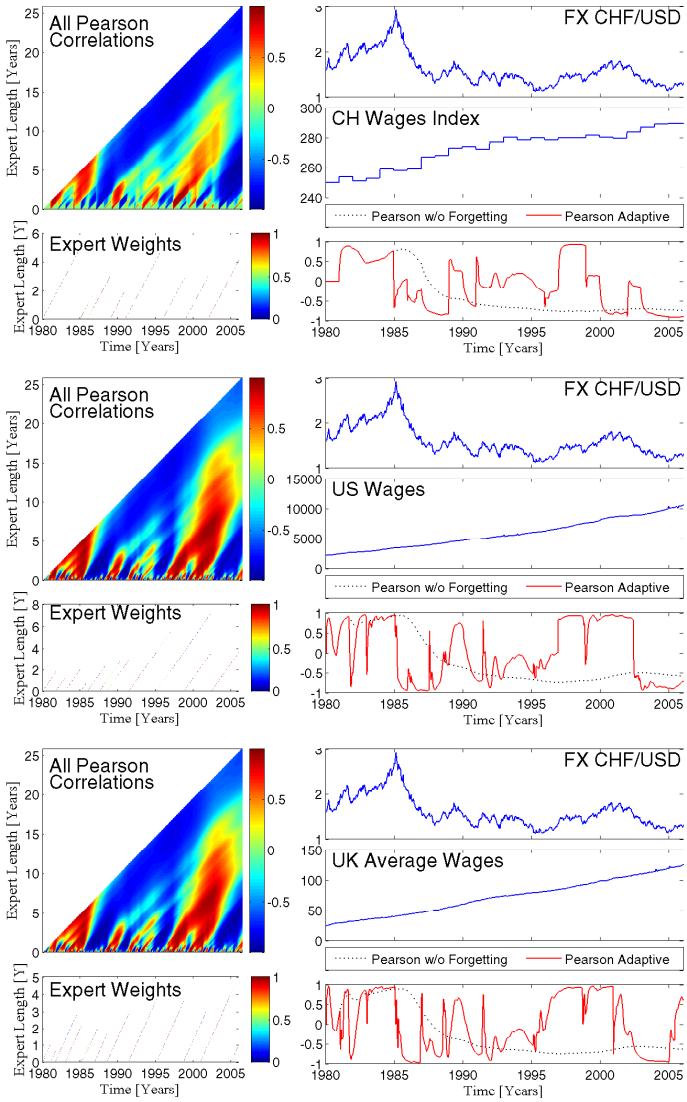


Figure 6.23: Wages.

The wages of the US and UK are continuously increasing in contrast to the Swiss wages which remain at the same level between 1993 and 2002. This behavior can also be recognized in the adaptive correlation curves.

6.3.14 Consumer Confidence Index

The Consumer Confidence Index (CCI) is calculated and published by the Conference Board. The latest CCI is published in the Board's monthly Consumer Confidence Survey[3]. The Consumer Confidence Survey contains details on consumer attitudes and buying intentions. Data is available by age, income and region. The reference of the index is the starting year 1967, when the index was set to 100. The CCI for EU and CH are calculated in a similar way, but with different reference levels.

The Bloomberg classification is medium important.

The CCI curves of the EU and (more coarse-grained) CH region are very similar. The US CCI is also similar, but does not show that pronounced ups and downs.

The adaptive Pearson correlation reveals some more information. It shows a high positive over-all correlation between the exchange rate and the US CCI, whereas the CH CCI does not show such distinct behavior - there is little correlation. The EU CCI shows high positive and negative correlations staying over years at the same level.

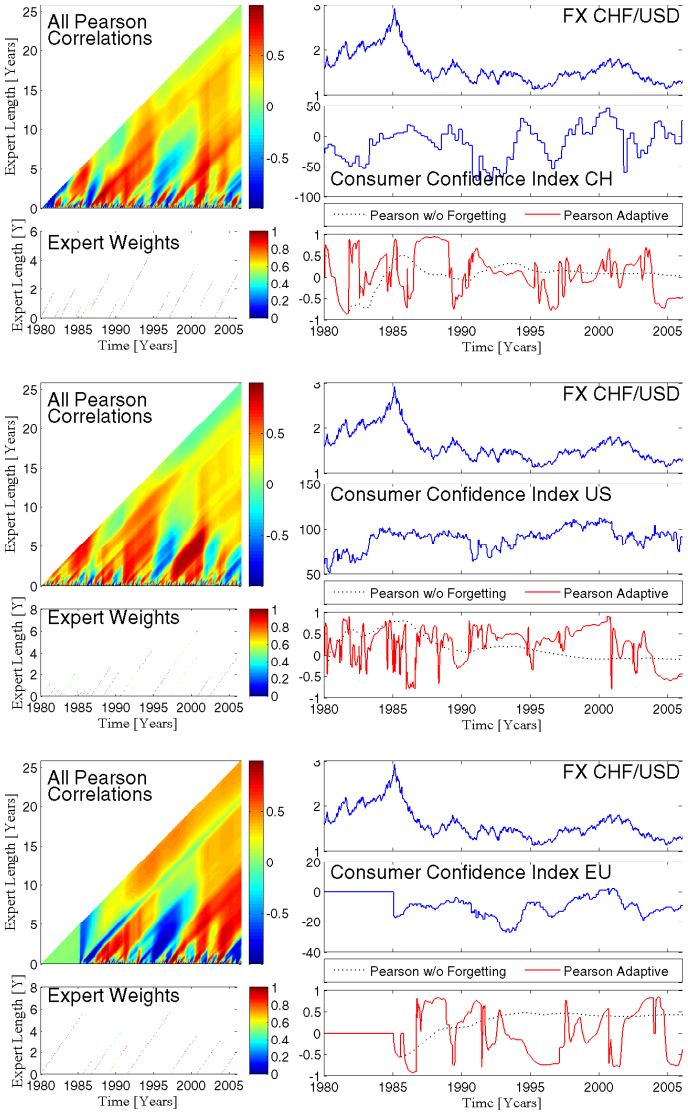


Figure 6.24: Consumer Confidence Index.

[3]The survey is posted on the Board's Web site www.conference-board.org/economics/consumerConfidence.cfm (October 6, 2007).

6.4 Discussion

All results show the advantage of using the adaptive Pearson correlation compared to the non-adaptive (without forgetting) correlation which turns out to be inadequate for this kind of problem.

According to the finance experts, they feel confirmed in their assumptions.On the one hand in observing recent situations and comparing it to the past. On the other hand as an early-warning system when something begins to change. Concerning the huge amount of variables of interest the experts like the overview on all correlations so they can focus directly on the hot spots. They are also able to compare the adaptive correlations with each other. The surface plot containing all correlations supports the expert in recognizing past patterns. Additionally, the comparison of the expert weights supports the experts in assessing different variables in parallel. So, even when the adaptive correlation values are different, the underlying drift can be similar and, thus, there might be a non-obvious relationship.

We can also identify global concept drifts, i.e. drifts that occur in most of the variables presented above. As an example we demonstrate two of such global effects on the relationship between FX CHF/USD and the oil price in Figure 6.25.

When looking at the two surface plots on the left of Figure 6.25 we can identify such a drift at the beginning of the year 1997. The upper surface plot shows the change in all possible Pearson correlations. A new structure emerges at 1997. This structure is also detected by the linear regression based indicator in the lower surface plot. Other relationships like the relationship to the Treasury bonds in Figure 6.9 or the currency swaps in Figure 6.5 show this even more distinct. Actually, we can associate the drift to a major event: the East Asian financial crisis. The East Asian financial crisis, also known as the East Asian currency crisis or as the IMF crisis, had a huge impact on currency exchange rates [Weisbrot, 2007]. Many nations learned from this crisis and quickly built up foreign exchange reserves which have also influence on the funding of Treasury bonds. This is exactly what we observe in our data.

We have also identified another kind of global effect. It is rather slowly appearing than abrupt compared to the effect of the East Asian financial crisis. We observed that by-and-by correlations tend to be more and more stable and the correlations are higher, too (see Figure 6.25). We identified the market penetration by information systems as possible reason. Information systems allow more participants to take part on the market. Also gathering of market figures and purchase orders can be processed electronically. Thus, the processes are faster, allow parallel processing and bridge company barriers - even more, automated trading systems can be installed. All these effects cause a more competitive market situation. A competitive market exhibits faster reaction times and faster price finding processes. This is exactly what we observe. At the end the correlations are higher due to the Pearson correlation that is sensitive to synchronous curve movements. Non-synchronous, i.e. delayed, movements would result in lower correlations and

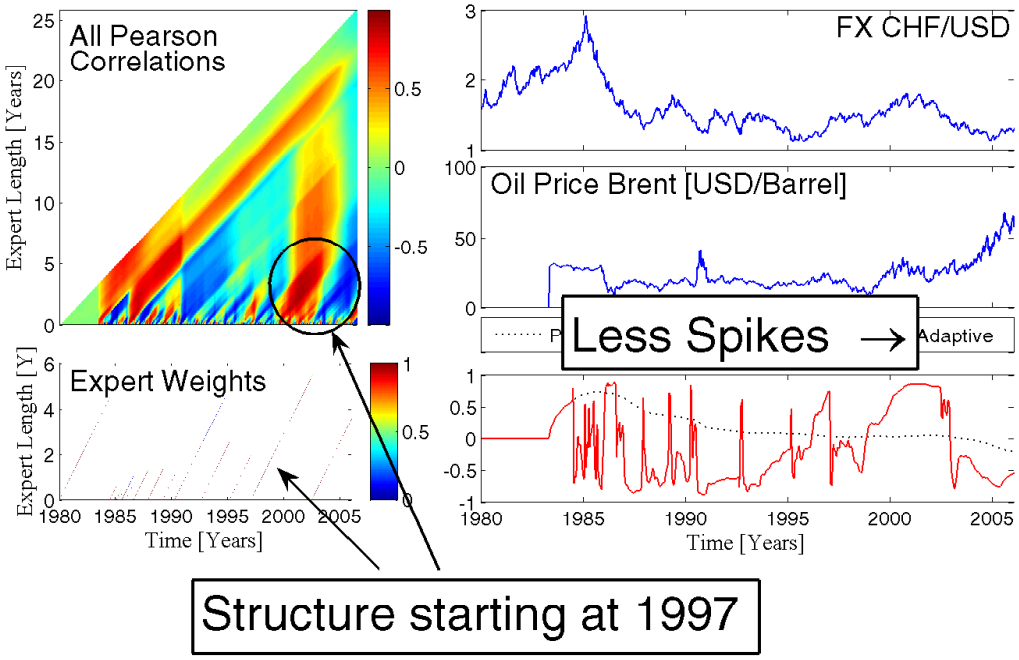


Figure 6.25: Exemplary global effect example.

spikes when the movements are parallel due to short-time overlapping. The assumption of the computer induced market change is supported by the correlations of the stock exchanges (Figure 6.13) which show high correlation for the longest time. Stock exchanges were the first trading platforms using new information systems. The influence of information systems can also be observed at the industrial production index (Fig. 6.3.10). The impact of information systems on production output starts in the mid-nineties.

The animated illustration of the regime drifts has been of high usefulness. It provides a more intuitive approach to the dynamics of the foreign exchange rate system. The expert is able to get a feeling about the fluctuations and the stability of the system and, thus, how reliable a variable might by.

This information combined with the knowledge of the experts augments the foreign exchange research and might have influence on trading strategies. But the investigation on this level is beyond the scope of this work and belongs to the daily practical business.

When dealing with correlations a fundamental question rises[4]: Is there a real correlation or a spurious correlation? Although finance experts feel confirmed in their assumptions when looking at the results, we have a look at an synthetic example where spurious correlations appear. For this purpose we have a closer look at the comparison of two random walk curves. Figure 6.26 is the illustration of such a scenario. Both random walk curves have been calculated by applying the

[4]Recall Chapter 2.1.5, p.16, bullet point 3: "The correlation is coincidental. The two events occur at the same time, they have no simple relationship to each other besides the fact that they are occurring at the same time."

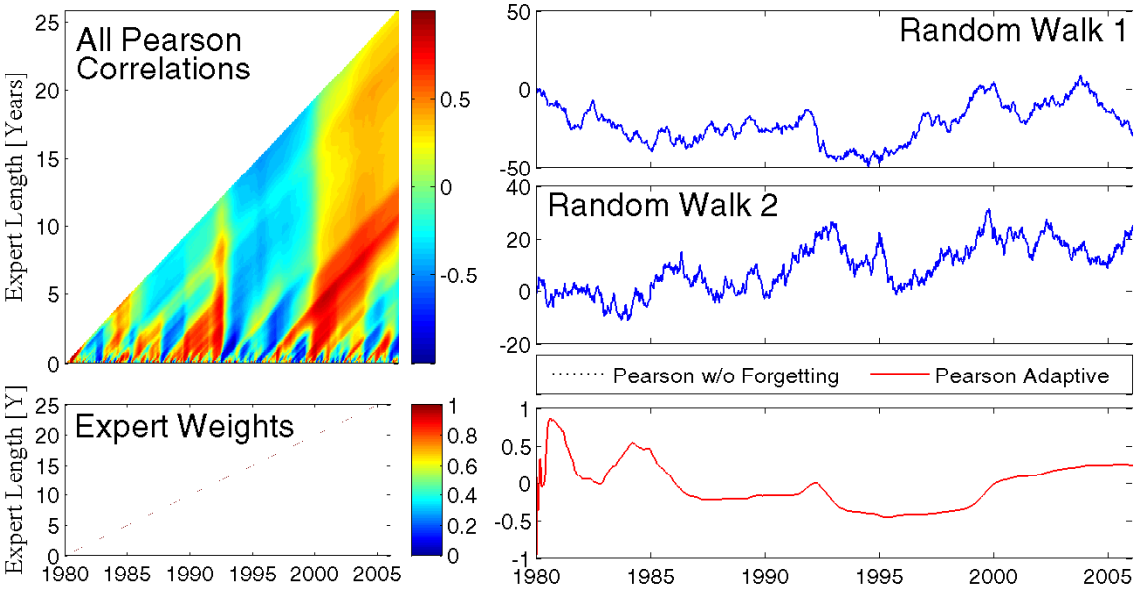


Figure 6.26: Computation for a random walk scenario.

formula $X(t+1) = X(t) + \Phi$, where the starting point $X(0) = 0$ and Φ is a random number from an uniform distribution between -1 and $+1$. Theoretically, the global overall correlation for infinite large and independent random walk curves tends to zero. Naturally, random walk curves move parallel in some intervals resulting in local non-zero correlations. The surface plot in the top left corner of Figure 6.26 shows such regions. In this case our regime handling method does not identify any drift and thus, does not collapse its window size[5]. So, our method does not fare worse than any general non-adaptive one. It will, however, detect random patterns, just like the non-adaptive ones. The correlation approaches the zero-value with increasing time series length as expected.

[5] All paramaters are the same as used for the finance calculations.

7

Limitations and Future Work

In this section we list identified limitations of our work. All limitations point to possible future research. Here, we distinguish between two factors limiting the validity of our study: the internal and external validity.

7.1 Internal Validity

Internal validity is the extent to which a study properly measures what it is meant to.

First, we examine some of our initial assumptions. We assumed simultaneous correlations. In the finance domain running ahead variables are known which are not covered by our correlation determination methods. Techniques like time warping [Berndt and Clifford, 1994] exist, but there is still research potential on time-shifted correlations. One of our findings was that the variables are getting more and more simultaneous. So, there is a trend towards simultaneous behavior which is covered by our measures.

The other assumption was the limitation to one-to-one relationships between all variables. From experience in the finance field we know that this assumption is reasonable. Theoretically, other effects might exist, for example, many-to-one relationships or self-reinforcing systems. The extension to such problems might contribute some value for finance models, but it has not been the focus of our research.

Furthermore, we have a look at the theoretical justification of applied techniques. Concept drift and feature selection are mainly empirical scientific disciplines. Even though we designed our experimental setup to ensure internal validity, our background is of empirical nature. Therefore, there is still potential for research on theoretical aspects.

Such research could focus on the interaction between dataset and algorithm. How fast does an algorithm approach the correct hypothesis depending on the nature of the dataset? The background is the decision at which point of time an old classifier should be substituted by a new one. This is of importance in the case of slowly and marginal drifting concepts. Especially, when

combining two different algorithms as seen in approach II. The indicator might be discarded, but the executing algorithm is not able to catch up with the current concept (or vice versa).

Theoretical work in this fields like the Vapnik-Chervonenkis VC dimension does not provide information about the learning dynamics [Vapnik and Chervonenkis, 1971]. Research on learning dynamics like active learning [Angluin, 1988] does not sufficiently cover theoretical investigations on the dataset / algorithm interaction – except for fastening the learning process depending on the dataset topology.

7.2 External Validity

External validity is the extent to which the results of a study can be generalized.

So, the assessment can be enhanced to more synthetic and real datasets. We could also consider more algorithms for correlation determination, for concept drift handling and for the ordinalization of the non-ordinal correlation values. Nevertheless, we are convinced that our study sufficiently ensures external validity.

8

Conclusions

In this work we exposed an actual problem addressed by finance researchers and traders. Then, we formalized this problem in terms of data mining methods. Specifically, we broke down the problem to the fields of feature selection and concept drift.

We tackled the problem by combining these two fields using two different approaches. The two approaches enlightened the problem from two different perspectives. On the one hand from the feature selection perspective and on the other hand from the concept drift perspective.

After implementing the two approaches we assessed them on two different synthetic datasets and a real world dataset. The results showed that both approaches are suitable for regime drift problems - even under noisy conditions. We compared the two approaches and selected the second approach because of its superior behavior concerning computational complexity.

Finally, we applied the selected method on real finance data. The results on the exchange rate data allow drawing conclusions that are consistent with real events. Even more, finance experts are able to gain more insight on the regime drift problem.

9

Acknowledgements

First of all, I like to thank my supervisor Prof. Avi Bernstein, Ph. D. for giving me the opportunity to pursue my research.

A big thank you goes to Prof. Dr. Marc Paolella for being the second reviewer which has been very important for me in the last months of my work. I also like to thank Prof. Dr. Gerhard Schwabe and Prof. Dr. Harald Gall for supporting me during my time at the department.

A special thank you to the students who wrote a diploma and/or bachelor thesis under my supervision (listing in alphabetical order).

Adrian Bachmann
Jens Balkausky
Domenic Benz
Simon Besmer
Simon Bleher
Robin Bucciarelli
Sandro Buccuzzo
Christoph Bürki
Stefan Christiani
Martin Constam
Michael Dänzer
Frédéric de Simoni
Manuel Donner
Patrice Egger
Marc Eichenberger
Joachim Fornallaz
Mathias Graf
Sinja Helfenstein
Andreas Huber
Urban Kägi
Lukas Kern
Matthias Linherr
Jonas Luell
Erika Mayer-Sommer
Andrej Nekrassov
Michael Polli
Esther Rölli
Giachem Schucan
Jan Sohnrey
Iwan Stierli
Daniel Suter
Roger Trösch
Stephan Tschanz
Benjamin Voigt
Alen Zurfluh

A thank you my co-assistants for fruitful research discussions and leisure time. Particularly, Jonas Luell, Jiwen Li, Michael Dänzer, Jayalath Ekanayake, Adrian Bachmann, Esther Kaufmann, Jonas Tappolet, Christoph Kiefer, Katharina Reinecke, and Stephanie Hauske.

I like to thank the IFI staff. Particularly, Enrico Solcá, Beat Rageth, Eveline Suter Schwarz, Rosa Frei, Samira Ludi, Zehra Kilit, and Lotti Kündig for their outstanding service.

Last, but not least, I thank Stanislav Yevgrafovich Petrov for demonstrating that a system is only as intelligent as its operator.

A

Appendix

A.1 Predictive Modeling Algorithms

Predictive modeling embraces classification and regression [Duda et al., 2000]. In this section we present the base algorithms we use throughout this work. The algorithms are the most used algorithms in machine learning as the ICDM 06 Panel[1] confirms. In the ICDM 06 Panel the classifiers introduced below are all amongst the "Top 10 Algorithms in Data Mining". Our additional criterion for the choice is the complementary structure of the different algorithm designs.

A.1.1 Classifiers

In this work four kinds of classifiers are applied. All classifiers below are based on the MATLAB implementation available from [Stork and Yom-Tov, 2004] except the Naïve Bayes Algorithm which has been implemented from scratch.

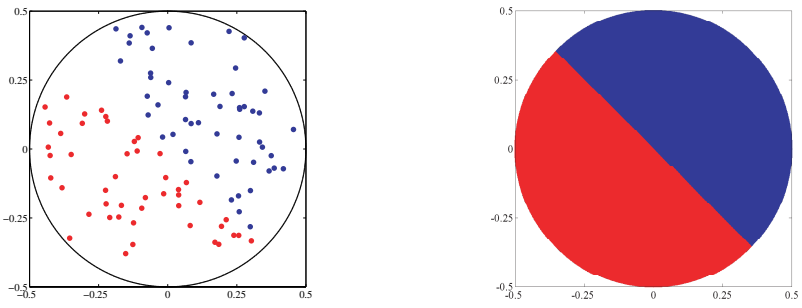


Figure A.1: Training and reference set for the visualization of the classifiers on the "sphere through plane" dataset.

[1] International Conference on Data Mining ICDM, http://www.cs.uvm.edu/~icdm/algorithms/ICDM06-Panel.pdf (December 21, 2006)

We provide visualizations for more insight into the classifier's learning mechanisms. The visualizations are based on the models learned on the training data illustrated at the left Figure A.1. The Figure on the right shows the perfect class separation for this two-class problem according the underlying data generating mechanism (concept). The concept definition is taken from the "plane intersects sphere" dataset in Section 2.2.4.

Decision Tree

We have chosen to apply Quinlan's C4.5 [Quinlan, 1993] as decision tree implementation. This tree uses information content as splitting criterion for continuous features and a histogram for discrete features. We modified the MATLAB C4.5 implementation in order to be able to handle special cases like one-dimensional patterns and to avoid infinite recursions. The decision tree parameter "confidence level" is the maximum error percentage at a node that will prevent it from further splitting.

Throughout this work we make use of two decision tree alternatives, one with the default confidence level of 25% and the other with a confidence level of 0%. A confidence level of 0% corresponds to a full decision tree without pruning. Typically, ensembles of unpruned trees perform better compared to ensembles of pruned trees [Sollich and Krogh, 1996, Street and Kim, 2001].

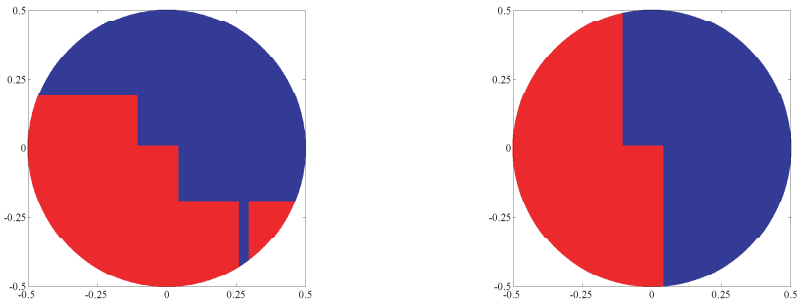


Figure A.2: Illustration for the decision tree models.

The two illustrations in Figure A.2 show the performance of the two decision trees on the "sphere through plane" dataset. The upper Figure illustrates the model of the unpruned decision tree. Since the splitting nodes are parallel to the input features the decision boundaries are vertical and horizontal, approximately close to the diagonal reference boundary. The lower Figure shows the pruned decision tree which only considers the two topmost decision node levels.

k-Nearest Neighbor

The k-nearest neighbor algorithm KNN is a type of instance-based learning [Dasarathy, 1990]. KNN is a method for classifying instances based on the closest k training examples in the feature space and, thus, is able to represent very complex models.

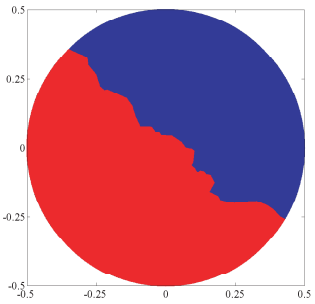


Figure A.3: Illustration for the KNN model.

Throughout this work we set the number of nearest neighbors to $k = 7$, unless the total number of instances dropped below 10 instances, where we reduced k accordingly. The applied metrics is the Euclidean distance.

The Figure A.3 represents the KNN model which represents the target concept very accurately.

Support Vector Machine

The Support Vector Machine SVM algorithm works in two stages [Boser et al., 1992]. In the first stage, the algorithm transforms the input data by a kernel function. In the second stage, the algorithm inserts a linear separating hyperplane. Throughout this work a Radial Basis Function RBF kernel function was used with its default Gaussian width of 0.05. A simple farthest-margin Perceptron solver is used to find the linear separating hyperplane.

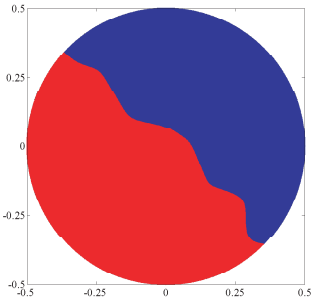


Figure A.4: Illustration for the SVM model.

The MATLAB implementation is only suitable for classification problems, although there exist SVMs for regression problems [Drucker et al., 1997]. We modified the algorithm in order to be able to handle special cases like one-dimensional patterns or situations where no support vectors could be found.

The Figure A.4 shows how well the SVM model represents the target concept. The boundary is smooth due to the Gaussian kernel function.

Naïve Bayes

The probabilistic classifier Naïve Bayes is based on the Bayes' theorem. The prefix "Naïve" stands for the assumption of the input features being independent of each other. [Zhang, 2004] assesses and gives some more insight in the Naïve Bayes algorithm.

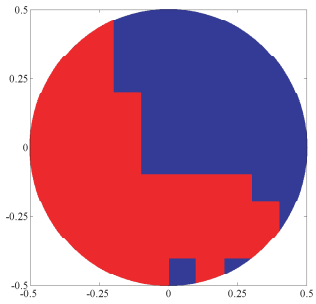


Figure A.5: Illustration for the Naïve Bayes model.

This implementation of the algorithm makes use of the Laplace estimation to avoid estimations biased by zero-probabilities. The Laplace values used throughout all calculations has been set to 0.1.
For the "plane intersects sphere" dataset all input features have been discretized into 10 equal-sized chunks. Figure A.5 shows how well the Naïve Bayes classifier copes with this kind of classification problem even though the features are assumed to be independent.

A.1.2 Regression

To perform regression we used the built-in MATLAB function `polyfit`. This function returns the polynomial coefficients of a n-degree polynomial fitting the data in a least squares sense. For this work we limited the problem to linear regression (n=1).

A.2 DWM Algorithm for Regression Problems

The state-of-the-art ensemble weighting algorithm is the Dynamic Weighted Majority DWM algorithm introduced by [Wang et al., 2003]. This algorithm is designed for classification problems. So, we adjusted the DWM algorithm since we are not aware of any other similar algorithm dealing with regression.

Table A.1 shows the original DWM algorithm on the left side and the adjusted algorithm on the right side. Both algorithms still look similar except for the following adjustments.

1. The prediction values Λ and λ are continuous because we are dealing with a regression problem. The λ results are stored for each expert for the subsequent calculation of the global prediction Λ, therefore, we use the annotation of λ_j.

2. The error ξ is compared to the threshold ϑ to eliminate non-acceptable ensemble experts. This is in contrast to the binary comparison of the prediction hitting the correct class.

3. The global prediction Λ is not chosen by taking the prediction of dominant expert as seen at the original algorithm. We calculate the overall prediction by adding up the expert's weighted predictions, i.e., we take the weighted average prediction value. The result is a more robust prediction. In the original algorithm we are dealing with class predictions, where we can not define averaged predictions unless the classes correspond to ordered numbers.

For the linear regression algorithm we define the error calculation function as the difference of the predicted value with respect of the real value $\mathrm{CalcError}\,(a, b) := |a - b|$.

The main difference between the two algorithms is the choice of the threshold parameter ϑ which determines the adaptivity of the ensemble towards concept drifts. We found the values depend on the applied domain, but they are quite robust under variations. So, for the "Plane through sphere" and the "meteorology" dataset we used a value of 0.5. For the "Stagger" dataset we applied a value of 0.2. The results on the "Stagger" dataset do not change much by taking a value of 0.5 – the algorithm would be a little bit less adaptive, but more robust under noise. For the finance dataset we used a value of 0.1. Before performing these calculations we normalized all data streams to a range between 0 and 1.

For the "Stagger" and "meteorology" dataset the period p between the reconsiderations is equal to 1. We set period p to 5 for the "Plane through sphere" and "finance" datasets because of their large number of instances. The other parameters have been chosen to be the same as recommended by the authors of the DWM algorithm. The threshold for deleting experts is $\theta = 0.01$ and the factor for decreasing weights is $\beta = 0.5$.

We also make use of the adjusted DWM algorithm at the method presented in the approach I on page 31, where we generated an ordinal measure based on non-ordinal values (see Fig. 3.4).

To transfer the new problem to a well-known problem we interpret the generated ordinal value as an error. Hence, we replace the error function $\mathrm{CalcError}(a, b)$ by the ordinal measure obtained for the corresponding ensemble expert. The λ predictions are replaced by the corresponding correlation values and the everything else remains the same. We kept the threshold values ϑ the same as defined for the regression problem.

$\{\vec{x}, y\}_1^n$: training data, feature vector and target
β : factor for decreasing weights, $0 \leq \beta \prec 1$
$c \in \mathbb{N}^*$: number of classes
$\{e, w\}_1^m$: set of experts and their weights
Λ, λ: global and local predictions
$\vec{\sigma} \in \mathbb{R}^c$: sum of weighted predictions for each class
θ: threshold for deleting experts
p: period between expert removal, creation, and weight update
ξ: error of prediction
ϑ: threshold for prediction deviation

DWM Classification

```
for i = 1, ..., n
    σ⃗ ← 0
    for j = 1, ..., m
        λ = Predict (e_j, x⃗_i)

        if (λ ≠ y_i and i mod p = 0)
            w_j ← β w_j
        σ_λ ← σ_λ + w_j
    end;
    Λ = argmax_λ σ_λ
    if(i mod p = 0)
        w ← NormalizeWeights(w)
        {e, w} ← DeleteExperts(e, w, θ)

        if(Λ ≠ y_i)
            m ← m + 1
            e_m ← CreateNewExpert()
            w_m ← 1
        end;
    end;
    for j = 1, ..., m
        e_j ← Train(e_j, x⃗_i)
    output Λ
end;
end.
```

DWM Regression

```
for i = 1, ..., n
    λ⃗ ← 0
    for j = 1, ..., m
        λ_j = Predict (e_j, x⃗_i)
        ξ = CalcError (λ_j, y_i)
        if (ξ > ϑ and i mod p = 0)
            w_j ← β w_j

    end;
    Λ = (1/w̄) Σ_j w_j λ_j, where w̄ = Σ_j w_j
    if(i mod p = 0)
        w ← NormalizeWeights(w)
        {e, w} ← DeleteExperts(e, w, θ)
        ξ = CalcError (Λ, y_i)
        if(ξ > ϑ)
            m ← m + 1
            e_m ← CreateNewExpert()
            w_m ← 1
        end;
    end;
    for j = 1, ..., m
        e_j ← Train(e_j, x⃗_i)
    output Λ
end;
end.
```

Table A.1: Pseudo-code for the DWM and adjusted DWM algorithm.

A.3 Application of Approach I on Classification Problems

In this section we benchmark the method presented as approach I in Chapter 3 in a well-known field: a classification task subjected to drifting concepts. The tests are performed on the two synthetic datasets "Stagger" and "plane through sphere" (see Section 2.2.4). The benchmark algorithm is the Dynamic Weighted Majority DWM algorithm[2] which has been designed for this kind of problems. As base algorithms we used the five classifiers presented in Appendix A.1.1.

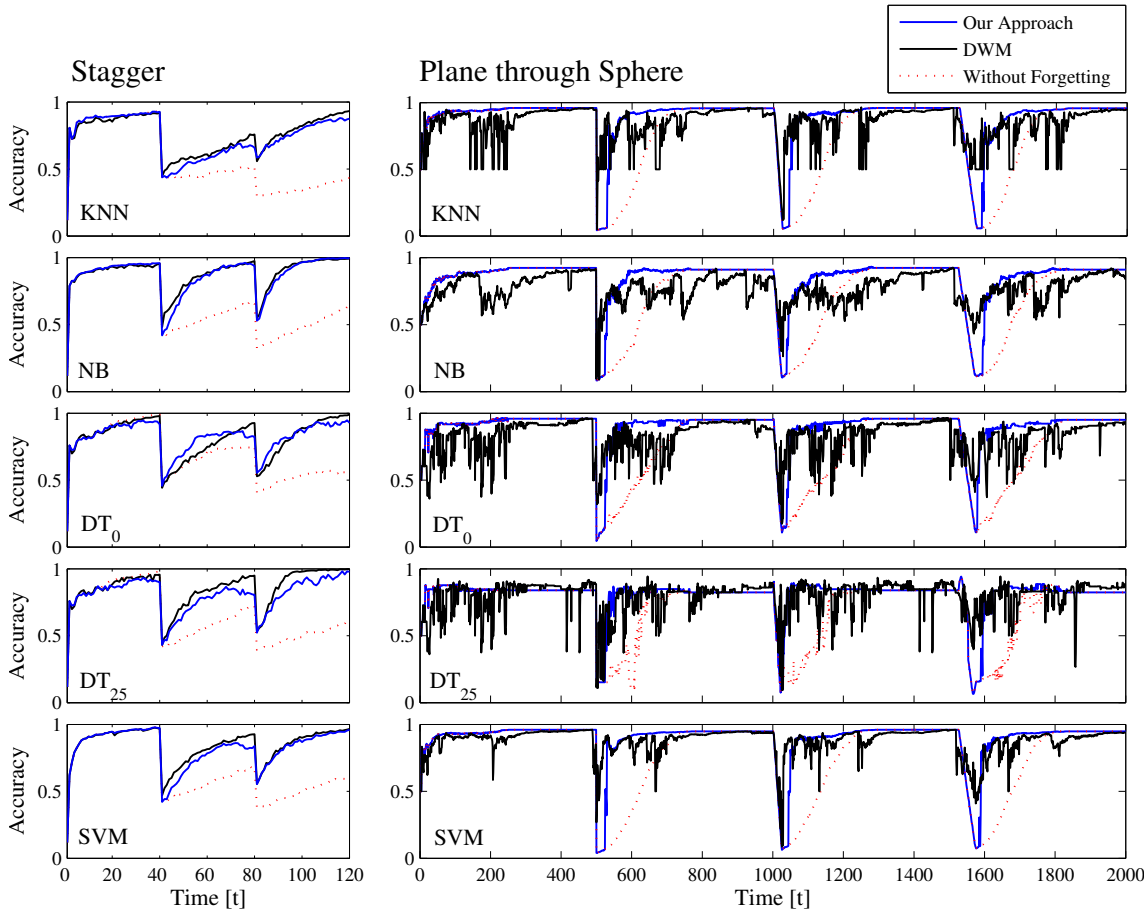


Figure A.6: Overview on the classification performance.

For the evaluation, our approach has been adjusted to be able to handle classification tasks. Instead of taking correlations as input values we now take accuracies. We are dealing with the accuracies as if they were correlations, i.e., we do not make use of the ordinal nature of the accuracies. Instead of combining correlations according to the identified drifts we combine the classifier predictions to get a final class prediction. So, the only adjustment we did is the rounding of the

[2]We used the same parameters as recommended by the authors [Wang et al., 2003].

final output values at the end to obtain whole-numbers corresponding to discrete class values[3].

The evaluation measure of the predictions is the accuracy. The accuracy is calculated by assessing the generated models on test sets. The evaluation has been conducted in the following way. On the "Stagger" dataset the classifiers are evaluated on test sets of 100 randomly generated instances of the current target which is presented to the learner at each time step. The assessment of the classifiers on the "plane through sphere" dataset is similar with the difference of presenting 10000 random instances to the learner.

Figure A.6 shows the resulting accuracies of both ensemble methods on both datasets. Additionally, there is a curve representing the performance of a non-adaptive algorithm. Both ensemble-based methods outperform the non-adaptive calculations. On the "Stagger" dataset the DWM benchmark is slightly better performing than the method of our approach. On the "plane through sphere" dataset the benchmark algorithm is more aggressive. This results in higher adaptivity - but less robustness - compared to the method of our approach.

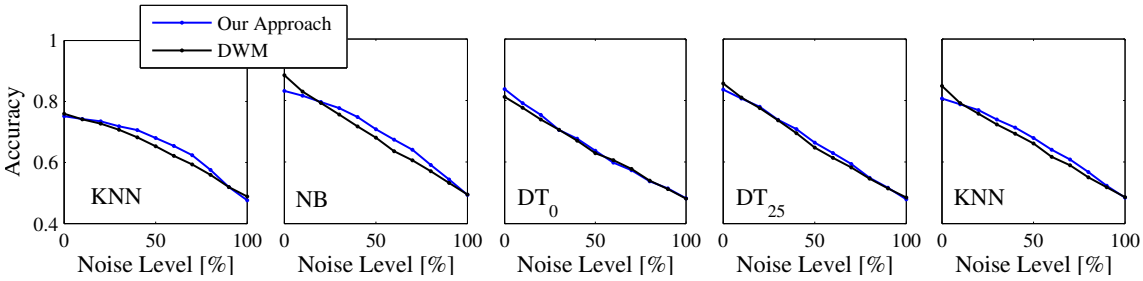


Figure A.7: Noise considerations for classification on "Stagger" dataset.

Figures A.7 and A.8 show the averaged accuracies on both datasets for different noise levels (up to 100%). The behavior is about the same for both datasets and all classifiers. The method of our approach is mostly more robust under noise influence. The reason is the ordinalization step of our method which contains smoothing that makes the predictions more robust towards noise.

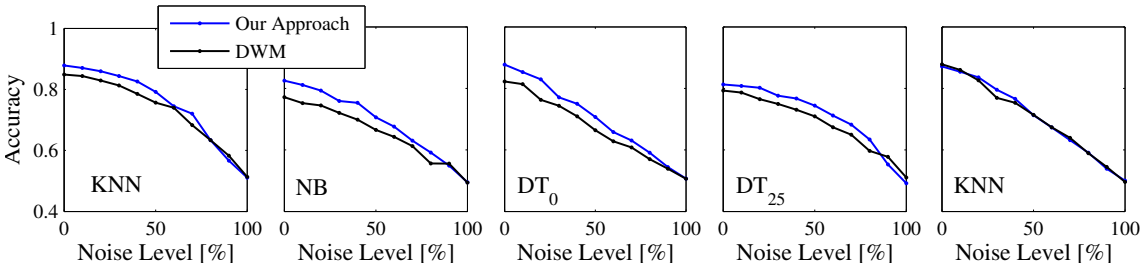


Figure A.8: Noise considerations for classification on "plane through sphere" dataset.

[3]This is allowed since we are dealing with two-class problems.

To conclude, both approaches perform on about the same level. This is surprising because the method of our approach does not have as much information about the problem compared to the original DWM algorithm. To achieve this performance we have to invest much more computational power (see Section 3.4.1, p. 42) than the DWM algorithm. So, we can make two statements. First, the method of our approach is suitable to detect drifting concepts which justifies its use for regime drift handling. Second, the results here suggest there could be constructed an even better performing and more robust ensemble classification method by combining these two approaches.

A.4 External Indication

Concept drift methods can be decomposed the same way. There is a drift indicating algorithm (indicator) and an algorithm (executor) doing some task based on the information obtained from the indicator. For example, the indicator's outcome is a collapsing of the window size when faced to a concept drift. Then, the executor performs its calculations based on the collapsed window size.

In most cases the indicator and executor are based on the same algorithm (see Figure A.9a). In some cases the indicator would be very time consuming when based on the same algorithm as the executor. A solution approach is a less complex algorithm as indicator preceding the more complex executor (see Figure A.9b). We call this approach "external indicator". For example in the field of feature selection – not in the field of concept or regime drifts – [Bi et al., 2003] use a linear indicator to select the features and learn on these features a non-linear SVM model (executor) which produces good results. This study of mixing two algorithm categories promises also good results for our approach; even though our application field is of dynamic nature.

The research question is whether an indicator algorithm can be substituted by another indicator algorithm or not – without any considerable loss of performance. This is a non-trivial question. An indicator algorithm might be too lazy in recognizing a concept drift and the executor will still use outdated data. Or the opposite, an algorithm might be too aggressive and useful information for the executor's model is thrown away. Though the VC-dimension theoretically describes the size of an algorithm's hyperspace (model building capacity), the VC-dimension provides no statement about how fast a certain dataset can be approached by the algorithm's hypotheses. This is the reason why we provide an empirical answer to this question in the two sections below.

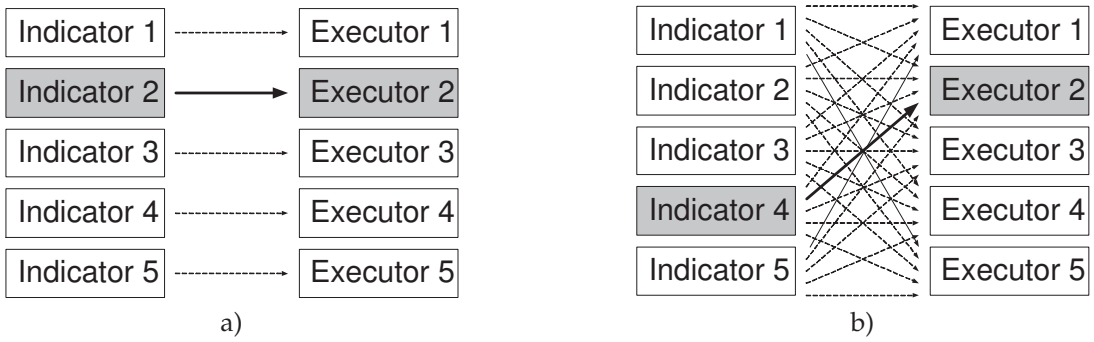


Figure A.9: Illustration of the external indicator setup.

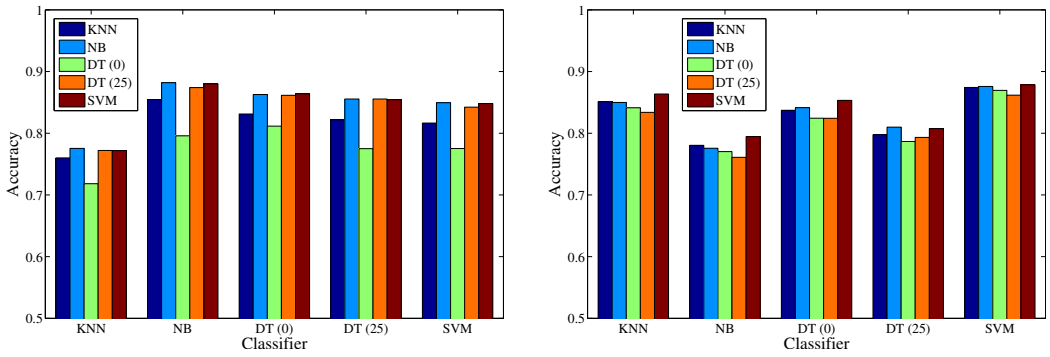
First, we have a look at a classification problem subjected to concept drifts. Then, we have a look at the regime drift problem using the method discussed in Approach II (see Section 4, p. 45).

The results below are presented as bar plot figures (see Fig. A.10 and A.11) that are composed as follows. The horizontal axis is grouped by the executing algorithms; each bundle contains the assessment values for all indicator combinations. The performance is illustrated by the value on

the y-axis.

A.4.1 Cross-Indication for Drifting Classification Problem

In this section we assess the classification task under drifting concepts using the five classifiers presented in Appendix A.1.1. We examine the influence of substituting an indicator classifier by the other classifiers. Figure A.10 shows the average accuracies reached by the five different indicator classifiers combined with all five executor classifiers. As ensemble selection method we use the Dynamic Weighted Majority DWM algorithm.



a) b)

Figure A.10: Cross-Indication assessment for a classification problem on a) the "Stagger" dataset and on b) the "plane intersects sphere" dataset

The result for both datasets is that the influence of the different indicators is marginal. Only the decision tree with confidence level 0 reveals it's shortcomings in dealing with the Stagger dataset[4]. So, indicating algorithms can be substituted by other algorithms to some extent. The impact of this finding is that we can perform the indicator step with less computational costs - say a Naïve Bayes classifier - and use a more time-consuming classifier for the final prediction like a Support Vector Machine.

A.4.2 External Indication for Drifting Regimes Problem

The regime drift problem following approach II consists of an indicator algorithm defining the drift handling and an executing algorithm which is a correlation finding algorithm. The concept drift detection has been performed with the DWM algorithm based on the six algorithms introduced in Appendix A.1. Then, the resulting ensemble expert weights are applied to determine the correlation values obtained from the feature ranking methods (see Section 2.1.4).

[4]The classifier's performance is very weak for short window sizes, thus, the ensemble (indicator) does not expand to larger and more robust experts

The performance of the regime calculations are reported in δ^2. δ^2 is the deviation between the calculated and the reference regime values. The results are comparable to the results of the previous section. The different overall values of the bundles are due to the different correlation (regime) value scales of the correlation determination methods. The only exceptions are the decision tree based methods on the Stagger dataset due to the short-coming of the decision trees on this kind of data[4]. The short-coming of the decision tree based indicator (confidence level 0) is apparent for the Pearson correlation.

The conclusion is that the overall indicator algorithms can be substituted to some extent. Maybe it's good to perform a trial before relying on one single indicator. So, the results are consistent with the results obtained for the pure classification problem in Section A.4.1.

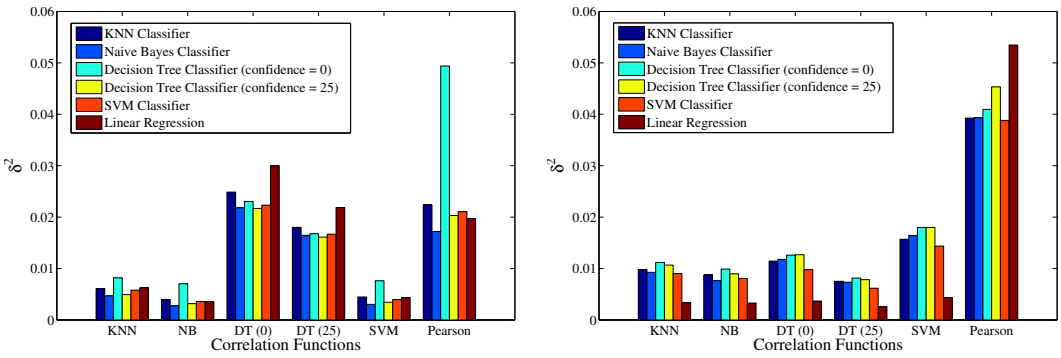


Figure A.11: External Indication assessment for a regime drift problem on a) the "Stagger" dataset and on b) the "plane intersects sphere" dataset

A.5 Noise Considerations

Figure A.12 shows the performance of both regime drift adaption approaches subjected to increasing noise levels. The assessment has been performed on the two synthetic datasets "Stagger" and "plane through sphere". Noise has been introduced by random switching of the target labels. As measure we use the δ^2 value which is the squared difference between the predicted correlation and the reference correlation averaged over all features. It is difficult to compare the δ^2 values since the absolute correlation values (predictions and reference) for the different wrapper and correlation methods are different. To allow a comparison between the different algorithms we normalized the δ^2 values by their δ^2 values at 100% ($\delta^2_{100\%}$). The sub figures in the rows show the assessment of the different approaches and the columns show the different datasets.

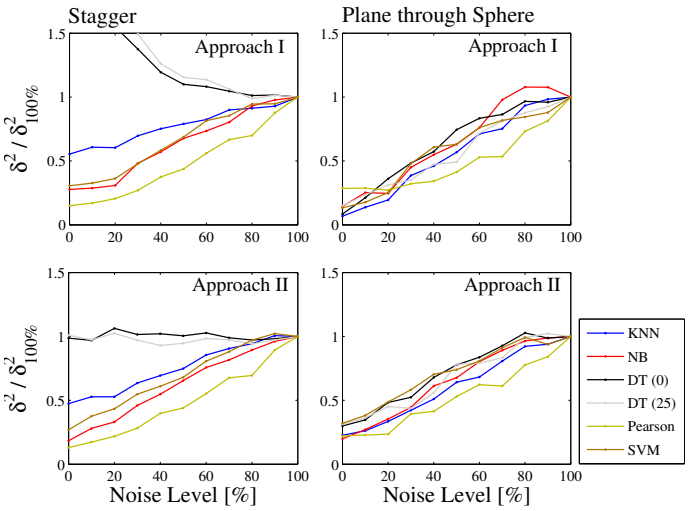


Figure A.12: Assessment of Approach I and Approach II on the two synthetic datasets for different functions and increasing noise levels.

First, we have a look at the first row, the assessment of approach I. The decision tree based wrapper value predictions are very poor as discussed in Section 3.3.1. The performance under low noise is even worse than the prediction under 100% noise. This is because the selected ensemble experts are of short window size and reflect different models than the reference models obtained at large window sizes. This is due to the limitations of the decision tree to model the Stagger dataset on few instances. On the "plane through sphere" dataset all correlation functions show the same behavior.

Second, we have a look at the assessment of approach II. Here, the performance of the decision tree based predictions perform poor on the Stagger dataset, too. But the performance is not much worse than the performance at 100% noise. On the "plane through sphere" dataset all correlations, again, perform similar.

All except the decision tree based curves perform very well and show a continuous loss of predictive performance. The absence of sudden changes suggests a stable behavior under noise.

Bibliography

[Aggarwal et al., 2003] Aggarwal, C. C., Han, J., Wang, J., and Yu, P. S. (2003). A framework for clustering evolving data streams. In *VLDB 2003: Proceedings of the 29th international conference on Very large data bases*, pages 81–92. VLDB Endowment.

[Angluin, 1988] Angluin, D. (1988). Queries and concept learning. *Mach. Learn.*, 2(4):319–342.

[Barnard, 1982] Barnard, G. A. (1982). Causation. *Encyclopedia of statistical sciences*, 1:387–389.

[Bauer and Kohavi, 1999] Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Mach. Learn.*, 36(1-2):105–139.

[Becerra-Fernandez et al., 2002] Becerra-Fernandez, I., Zanakis, S. H., and Walczak, S. (2002). Knowledge discovery techniques for predicting country investment risk. *Comput. Ind. Eng.*, 43(4):787–800.

[Benford, 1938] Benford, F. (1938). The law of anomalous numbers. *j-PROC-AMER-PHIL-SOC*, 78(4):551–572.

[Berndt and Clifford, 1994] Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD Workshop*, pages 359–370.

[Bi et al., 2003] Bi, J., Bennett, K., Embrechts, M., Breneman, C., and Song, M. (2003). Dimensionality reduction via sparse support vector machines. *J. Mach. Learn. Res.*, 3:1229–1243.

[Bloomberg Financial Glossary, 2000] Bloomberg Financial Glossary (2000). Bloomberg financial glossary.

[BLS, 2004] BLS (March 18, 2004). How does the producer price index differ from the consumer price index?

[BLS Handbook of Methods, 2003] BLS Handbook of Methods (September 3, 2003). Bls handbook of methods.

[Blum and Langley, 1997] Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artif. Intell.*, 97(1-2):245–271.

[Boser et al., 1992] Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, New York, NY, USA. ACM Press.

[Box and Jenkins, 1994] Box, G. E. P. and Jenkins, G. M. (1994). *Time Series Analysis: Forecasting and Control*. Prentice Hall PTR, Upper Saddle River, NJ, USA.

[Breiman et al., 1984] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

[Brooks, 1991] Brooks, R. A. (1991). Intelligence without reason. In Myopoulos, J. and Reiter, R., editors, *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91)*, pages 569–595, Sydney, Australia. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA.

[Cash, 1979] Cash, W. (1979). Parameter estimation in astronomy through application of the likelihood ratio. *Astrophys. J.*, 228:939–947.

[CBS, 2007] CBS (2007). Triennial central bank survey of foreign exchange and derivatives market activity in april 2007 - preliminary global results. http://www.bis.org/publ/rpfx07.pdf (december 6, 2007).

[Celik and Karatepe, 2007] Celik, A. E. and Karatepe, Y. (2007). Evaluating and forecasting banking crises through neural network models: An application for turkish banking sector. *Expert Syst. Appl.*, 33(4):809–815.

[Chu et al., 2004] Chu, F., Wang, Y., and Zaniolo, C. (2004). An adaptive learning approach for noisy data streams. In *ICDM*, pages 351–354.

[Cun et al., 1990] Cun, Y. L., Denker, J. S., and Solla, S. A. (1990). Optimal brain damage. In *Advances in neural information processing systems 2*, pages 598–605, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Dasarathy, 1990] Dasarathy, B. V. (1990). *Nearest Neighbor: Pattern Classification Techniques (Nn Norms : Nn Pattern Classification Techniques)*. Los Alamitos: IEEE Computer Society Press.

[Dash and Liu, 2000] Dash, M. and Liu, H. (2000). Feature selection for clustering. In *PADKK '00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, pages 110–121, London, UK. Springer-Verlag.

[Dasu and Johnson, 2003] Dasu, T. and Johnson, T. (2003). *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons, Inc., New York, NY, USA.

[Dhillon et al., 2003] Dhillon, I. S., Mallela, S., and Kumar, R. (2003). A divisive information theoretic feature clustering algorithm for text classification. *J. Mach. Learn. Res.*, 3:1265–1287.

[Dietterich, 2000] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *MCS '00: Proceedings of the First International Workshop on Multiple Classifier Systems*, pages 1–15, London, UK. Springer-Verlag.

[Drucker et al., 1997] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik, V. (1997). Support vector regression machines. In Mozer, M. C., Jordan, M. I., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, volume 9, page 155. The MIT Press.

[Duch et al., 2004] Duch, W., Wieczorek, T., Biesiada, J., and Blachnik, M. (2004). Comparison of feature ranking methods based on information entropy. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2, pages 1415–1419. IEEE.

[Duda et al., 2000] Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification (2nd Edition)*. Wiley-Interscience.

[Džeroski, 2000] Džeroski, S. (2000). Data mining in a nutshell. In *Relational Data Mining*, pages 3–27, New York, NY, USA. Springer-Verlag New York, Inc.

[Encyclopædia Britannica, 2007] Encyclopædia Britannica (2007). *Encyclopædia Britannica*. Encyclopædia Britannica, Inc.

[Fama, 1970] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417.

[Fan, 2004] Fan, W. (2004). Systematic data selection to mine concept-drifting data streams. In *KDD '04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 128–137. ACM Press.

[Forman, 2003] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3:1289–1305.

[Galindo and Tamayo, 2000] Galindo, J. and Tamayo, P. (2000). Credit risk assessment using statistical and machinelearning: Basic methodology and risk modeling applications. *Comput. Econ.*, 15(1-2):107–143.

[Gately and Gately, 1995] Gately, E. and Gately, E. (1995). *Neural Networks for Financial Forecasting*. John Wiley & Sons, Inc., New York, NY, USA.

[Guyon and Elisseeff, 2003] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.

[Guyon et al., 2006] Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A. (2006). *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[Hall, 1998] Hall, M. (1998). Correlation-based feature selection for machine learning.

[Hall and Holmes, 2003] Hall, M. A. and Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(6):1437–1447.

[Harries and Horn, 1995] Harries, M. and Horn, K. (1995). Detecting concept drift in financial time series prediction using symbolic machine learning.

[Harries et al., 1998] Harries, M. B., Sammut, C., and Horn, K. (1998). Extracting hidden context. *Machine Learning*, 32(2):101–126.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning*. Springer.

[Helmbold and Long, 1994] Helmbold, D. P. and Long, P. M. (1994). Tracking drifting concepts by minimizing disagreements. *Mach. Learn.*, 14:27–45.

[Herbster and Warmuth, 1998] Herbster, M. and Warmuth, M. K. (1998). Tracking the best regressor. In *COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory*, pages 24–31, New York, NY, USA. ACM Press.

[Herbster and Warmuth, 2001] Herbster, M. and Warmuth, M. K. (2001). Tracking the best linear predictor. *J. Mach. Learn. Res.*, 1:281–309.

[Holland, 1986] Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

[Huang et al., 2004] Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., and Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decis. Support Syst.*, 37(4):543–558.

[Hulten et al., 2001] Hulten, G., Spencer, L., and Domingos, P. (2001). Mining time-changing data streams. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 97–106, San Francisco, CA. ACM Press.

[Jolliffe, 2002] Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

[Kenneth, 2003] Kenneth, S. (2003). Learning concept drift with a committee of decision trees, Tech. Report UT-AI-TR-03-302.

[Kennon, 2007] Kennon, J. (2007). Investing for beginners.

[Kifer et al., 2004] Kifer, D., Ben-David, S., and Gehrke, J. (2004). Detecting change in data streams. In *VLDB*, pages 180–191.

[Kingdon, 1997] Kingdon, J. (1997). *Intelligent Systems and Financial Forecasting*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[Kirkland et al., 1998] Kirkland, J. D., Senator, T. E., Hayden, J. J., Dybala, T., Goldberg, H. G., and Shyr, P. (1998). The nasd regulation advanced detection system (ads). In *AAAI '98/IAAI '98: Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, pages 1055–1062, Menlo Park, CA, USA. American Association for Artificial Intelligence.

[Klinkenberg and Joachims, 2000] Klinkenberg, R. and Joachims, T. (2000). Detecting concept drift with support vector machines. In *ICML*, pages 487–494.

[Klinkenberg and Rüping, 2003] Klinkenberg, R. and Rüping, S. (2003). Concept drift and the importance of examples. In *Text Mining – Theoretical Aspects and Applications*, pages 55–77. Physica-Verlag, Berlin, Germany.

[Kohavi and John, 1997] Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artif. Intell.*, 97(1-2):273–324.

[Kolter and Maloof, 2003] Kolter, J. Z. and Maloof, M. A. (2003). Dynamic weighted majority: A new ensemble method for tracking concept drift. In *ICDM*, pages 123–130.

[Kovalerchuk and Vityaev, 2005] Kovalerchuk, B. and Vityaev, E. (2005). *Data Mining for Financial Applications*. Springer.

[Kuh et al., 1990] Kuh, A., Petsche, T., and Rivest, R. L. (1990). Learning time-varying concepts. In *NIPS-3: Proceedings of the 1990 conference on Advances in neural information processing systems 3*, pages 183–189. Morgan Kaufmann Publishers Inc.

[Kuncheva, 2004] Kuncheva, L. I. (2004). Classifier ensembles for changing environments. In *Multiple Classifier Systems*, pages 1–15.

[Lazarescu et al., 2004] Lazarescu, M. M., Venkatesh, S., and Bui, H. H. (2004). Using multiple windows to track concept drift. In *Intelligent Data Analysis*, volume 8, pages 29–59.

[Littlestone and Warmuth, 1994] Littlestone, N. and Warmuth, M. K. (1994). The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261.

[Liu and Yu, 2005] Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502.

[Lyapunov, 1992] Lyapunov, A. M. (1992). The general problem of the stability of motion (in russian, reprinted in english, taylor & francis, london, 1992). *Comm. Soc. Math. Kharkow*.

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw Hill.

[Morgan Guaranty, 1994] Morgan Guaranty (1994). Riskmetrics technical document 2nd edition.

[Nakhaeizadeh et al., 2002] Nakhaeizadeh, G., Steurer, E., and Bartlmae, K. (2002). *Banking and finance*. Oxford University Press, Inc., New York, NY, USA.

[Nasraoui et al., 2003] Nasraoui, O., Cardona-Uribe, C., and Rojas-Coronel, C. (2003). Tecnostreams: Tracking evolving clusters in noisy data streams with a scalable immune system learning model. *ICDM*.

[Nonaka and Takeuchi, 1995] Nonaka, I. and Takeuchi, H. (1995). *The Knowledge-Creating Company : How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press.

[Opitz, 1999] Opitz, D. (1999). Feature selection for ensembles. In *AAAI/IAAI*, pages 379–384.

[Opitz and Maclin, 1999] Opitz, D. and Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198.

[Pearl, 2000] Pearl, J. (2000). *Causality : Models, Reasoning, and Inference*. Cambridge University Press.

[Phua et al., 2005] Phua, C., Lee, V., Smith, K., and Gayler, R. (2005). A comprehensive survey of data mining-based fraud detection research. *Artificial Intelligence Review*.

[Pourahmadi, 2001] Pourahmadi, M. (2001). *Foundations of Time Series Analysis and Prediction Theory*. John Wiley & Sons Inc.

[Provost and Fawcett, 2001] Provost, F. and Fawcett, T. (2001). Robust classification for imprecise environments. *Mach. Learn.*, 42(3):203–231.

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

[Rahman et al., 2002] Rahman, R. M., Thulasiram, R. K., and Thulasiraman, P. (2002). Forecasting stock prices using neural networks on a beowulf cluster. In *IASTED PDCS*, pages 465–470.

[Riecken, 2000] Riecken, D. (2000). Introduction: personalized views of personalization. *Commun. ACM*, 43(8):26–28.

[Rozsypal and Kubat, 2005] Rozsypal, A. and Kubat, M. (2005). Association mining in time-varying domains. *Intell. Data Anal.*, 9(3):273–288.

[Russell and Norvig, 2003] Russell, S. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach (Second Edition)*. Prentice Hall.

[Schlimmer and Granger, 1986] Schlimmer, J. C. and Granger, R. H. (1986). Beyond incremental processing: Tracking concept drift. In *Proc. of AAAI-86*, pages 502–507, Philadelphia, PA.

[Senator, 2000] Senator, T. E. (2000). Ongoing management and application of discovered knowledge in a large regulatory organization: a case study of the use and impact of nasd regulation's advanced detection system (rads). In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–53, New York, NY, USA. ACM.

[SNB Glossary, 2007] SNB Glossary (October 12, 2007). Snb glossary.

[Sollich and Krogh, 1996] Sollich, P. and Krogh, A. (1996). Learning with ensembles: How overfitting can be useful. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems*, volume 8, pages 190–196. The MIT Press.

[Stork and Yom-Tov, 2004] Stork, D. G. and Yom-Tov, E. (2004). *Computer Manual in MATLAB to Accompany Pattern Classification, Second Edition*. Wiley-Interscience.

[Street and Kim, 2001] Street, W. N. and Kim, Y. (2001). A streaming ensemble algorithm (sea) for large-scale classification. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 377–382, New York, NY, USA. ACM Press.

[Tsymbal, 2004] Tsymbal, A. (2004). The problem of concept drift: definitions and related work.

[UBS Dictionary of Banking, 2007] UBS Dictionary of Banking (2007). Ubs dictionary of banking 2007 edition.

[Vapnik and Chervonenkis, 1971] Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280.

[Varian, 1972] Varian, H. R. (1972). Letter to the Editor: Benford's Law. *j-AMER-STAT*, 26(3):65–66.

[Vorburger and Bernstein, 2005] Vorburger, P. and Bernstein, A. (2005). Entropy-based detection of real and virtual concept shifts.

[Vorburger and Bernstein, 2006a] Vorburger, P. and Bernstein, A. (2006a). Entropy-based concept shift detection. In *ICDM*, pages 1113–1118.

[Vorburger and Bernstein, 2006b] Vorburger, P. and Bernstein, A. (2006b). Entropy-based concept shift detection. In *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, pages 1113–1118, Washington, DC, USA. IEEE Computer Society.

[Walczak, 2001] Walczak, S. (2001). An empirical analysis of data requirements for financial forecasting with neural networks. *J. Manage. Inf. Syst.*, 17(4):203–222.

[Wang et al., 2003] Wang, H., Fan, W., Yu, P. S., and Han, J. (2003). Mining concept-drifting data streams using ensemble classifiers. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 226–235. ACM Press.

[Weigend, 1997] Weigend, A. S. (1997). Data mining in finance: Report from the post-nncm-96 workshop on teaching computer intensive methods for financial modeling and data analysis. *Proceedings of the Fourth International Conference on Neural Networks in the Capital Markets NNCM-96*, pages 399–412.

[Weisbrot, 2007] Weisbrot, M. (7 Aug 2007). Ten years after: The lasting impact of the asian financial crisis. In *CEPR Reports*, Washington, DC, US. Center for Economic and Policy Research (CEPR).

[Weston et al., 2003] Weston, J., Elisseeff, A., Schölkopf, B., and Tipping, M. (2003). Use of the zero norm with linear models and kernel methods. *J. Mach. Learn. Res.*, 3:1439–1461.

[Widmer, 1996] Widmer, G. (1996). Recognition and exploitation of contextual Clues via incremental meta-learning. In *International Conference on Machine Learning*, pages 525–533.

[Widmer and Kubat, 1992] Widmer, G. and Kubat, M. (1992). Learning flexible concepts from streams of examples: Flora2. In *ECAI '92: Proceedings of the 10th European conference on Artificial intelligence*, pages 463–467, New York, NY, USA. John Wiley & Sons, Inc.

[Widmer and Kubat, 1993] Widmer, G. and Kubat, M. (1993). Effective learning in dynamic environments by explicit context tracking. In *ECML '93: Proceedings of the European Conference on Machine Learning*, pages 227–243, London, UK. Springer-Verlag.

[Widmer and Kubat, 1996] Widmer, G. and Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Mach. Learn.*, 23(1):69–101.

[Witten and Frank, 2001] Witten, I. and Frank, E. (2001). *Data Mining*. Hanser Verlag München.

[Yoon and Shahabi, 2006] Yoon, H. and Shahabi, C. (2006). Feature subset selection on multivariate time series with extremely large spatial features. In *ICDMW '06: Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops*, pages 337–342, Washington, DC, USA. IEEE Computer Society.

[Yoon and Yang, 2005] Yoon, H. and Yang, K. (2005). Feature subset selection and feature ranking for multivariate time series. *IEEE Transactions on Knowledge and Data Engineering*, 17(9):1186–1198. Member-Cyrus Shahabi.

[Zhang and Berardi, 2001] Zhang, G. P. and Berardi, V. L. (2001). Time series forecasting with neural network ensembles: An application for exchange rate prediction. *The Journal of the Operational Research Society*, 53(6):652–664.

[Zhang and Hu, 1998] Zhang, G. P. and Hu, M. (1998). Neural network forecasting of the british pound
us dollar exchange rate. *International Journal of Management Science*, 26(4):495–506.

[Zhang, 2004] Zhang, H. (2004). The optimality of naive bayes. In Barr, V. and Markov, Z., editors, *FLAIRS Conference*. AAAI Press.